Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Doubly robust and interpretable evaluation of customer retention targeting in subscription businesses

xiaobo Xing

xingxiaobosys@sina.com

Heilongjiang University Of Technology

Research Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11319085/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Doubly robust and interpretable evaluation of customer retention targeting in subscription businesses

Xiaobo Xing[1*]

[1*]Heilongjiang University of Technology, Jixi, Heilongjiang, China.

Corresponding author(s). E-mail(s): xingxiaobosys@sina.com;

Abstract

Customer retention targeting requires estimates of incremental response, yet robust average-effect estimation does not guarantee reliable customer rankings. We develop a reproducible evaluation framework combining cross-fitted doubly robust scores, interpretable uplift ranking, covariate diagnostics, and resource-constrained decisions. An internet-service cohort of 5,517 IBM Telco sample records provides an observational illustration; technical-support subscription is a service-state proxy rather than a documented intervention. On 1,655 held-out records, the doubly robust top-20% contrast is 12.52 percentage points, with a conditional interval of $[-5.43, 28.75]$, and paired comparisons do not establish superiority over either baseline. Thirty repeated partitions expose substantial ranking variability. Two hundred average-effect simulations verify conditional nuisance robustness. A further 600 simulations evaluate known conditional effects and decision loss under logistic, interaction, and weak-overlap mechanisms. Adding second-stage interactions reduces doubly robust conditional-effect error from 11.32 to 7.96 percentage points under the interaction mechanism, but increases it from 9.34 to 11.66 under weak overlap. These findings distinguish causal identification, nuisance robustness, and decision reliability. The framework supports scrutiny of variable timing, model specification, and targeting uncertainty without claiming demonstrated campaign effectiveness or financial returns.

Keywords: Customer retention, Doubly robust estimation, Uplift modelling, Policy evaluation, Interpretability, Subscription analytics

1 Introduction

Subscription businesses use customer analytics to decide where limited retention resources should be allocated. Telecommunications providers, digital services, and membership platforms face recurring decisions about support, offers, and customer engagement. Churn prediction supplies one input to these decisions by identifying customers likely to leave. Reviews document substantial progress in predictive algorithms, feature construction, and validation procedures [1]. However, predictive accuracy alone does not identify which customers would benefit from a particular action.

The distinction between risk and responsiveness changes the targeting objective. A customer may have high churn risk but little prospect of responding to an offer. Another customer may remain subscribed without receiving additional support. Targeting either customer can consume resources without generating incremental retention. Evidence on retention targeting therefore questions the practice of automatically prioritizing the highest-risk customers [2]. Profit-oriented

approaches further require incremental response to be evaluated alongside campaign costs and customer value [3].

Uplift modelling addresses the incremental-response problem by comparing predicted outcomes under treatment and control [4]. In observational records, this comparison also requires assumptions about treatment selection and covariate timing. Support subscriptions, contract terms, and billing attributes are often recorded contemporaneously. Their availability does not establish that they preceded an intervention. A flexible model can describe these records accurately while leaving the intervention effect unidentified.

Three additional issues affect the credibility of retention evaluations. First, a doubly robust estimator requires residual correction terms, rather than only a difference between outcome predictions. Second, uncertainty in a targeted subgroup can be appreciable even when an aggregate association is precise. Third, accumulated tenure does not automatically constitute a follow-up time measured from treatment assignment. These issues concern the validity of the evidence chain, rather than the visual quality of a performance comparison.

This study develops a transparent framework for evaluating retention targeting under these constraints. We combine cross-fitted doubly robust scores, interpretable uplift ranking, propensity diagnostics, held-out policy evaluation, and conditional uncertainty estimates. A resource-allocation formulation connects statistical scores to costs without assuming that financial inputs are available in the benchmark. A separate longitudinal formulation states the additional information required for survival-based evaluation.

The study asks whether favourable measured balance and nuisance robustness are sufficient to support reliable customer targeting. Our contribution is an evaluation design that separates identification, average-effect robustness, and the quality of the resulting decision. We combine a public-data illustration, covariate sensitivity, repeated training partitions, and semi-synthetic experiments with known effects. The latter directly compare conditional-effect error and oracle-relative decision loss, including a controlled change to the second-stage function class. This design exposes a practical tension: expanding the ranking model can improve nonlinear effect recovery while worsening performance under weak overlap. We evaluate established estimators rather than proposing a new identification theorem or claiming universally superior prediction.

2 Related work

2.1 Prediction, uplift, and customer value

Predictive churn models estimate the observed outcome distribution under the existing service environment. Uplift models instead approximate differences between treatment-specific outcome functions [4]. These objectives can produce different customer rankings, and neither ranking by itself establishes a profitable campaign. Lemmens and Gupta incorporate campaign costs and customer cash flows into retention targeting [3]. Their distinction motivates our separation of incremental retention, coverage, and monetary value.

Two-model learners fit separate response models for treated and untreated customers. Meta-learners offer alternative ways to construct heterogeneous-effect estimates from these component models [5]. Causal forests provide a different approach to effect heterogeneity and inference [6]. Our implementation uses a two-model learner as a transparent baseline and a regularized regression of doubly robust scores as the uplift learner. This choice permits direct inspection of the score construction and its sensitivity to nuisance modelling.

2.2 Doubly robust learning and policy evaluation

Doubly robust heterogeneous-effect estimation combines outcome predictions with treatment-weighted residuals [7]. Its robustness property concerns nuisance estimation under an identified causal model. It does not eliminate omitted confounding, repair an ill-defined treatment, or establish the timing of covariates. Policy learning also distinguishes accurate effect estimation from the value of the action rule that uses those estimates [8]. We therefore evaluate rankings on held-out records and report paired differences between targeting rules.

2.3 Temporal retention and interpretability

Temporal customer churn methods can represent evolving interventions and event processes. Gao and colleagues develop a causal churn model using low-rank tensor structure and time-dependent hazard information [9]. Such methods require a longitudinal data structure that is absent from the 21-column benchmark used here. We retain survival analysis as a prospective extension and report observed tenure descriptively.

Interpretability also depends on the quantity being explained. An explanation of churn probability does not directly explain incremental treatment response. Variable-importance measures describe a fitted model and can depend on the available predictive alternatives [10]. Our grouped permutation analysis measures changes in predicted uplift, while subgroup estimates summarize doubly robust scores. Neither analysis assigns causal effects to individual covariates.

3 Evaluation framework

3.1 Observations, estimands, and identification

Let $O_i = (X_i, A_i, Y_i)$ denote a customer record. Here, X_i represents observed covariates, $A_i \in \{0,1\}$ denotes treatment, and $Y_i \in \{0,1\}$ indicates retention. Potential outcomes $Y_i(1)$ and $Y_i(0)$ describe retention under the two treatment states [11]. The conditional average treatment effect (CATE) is

$$\tau(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x\}. \tag{1}$$

This quantity averages over customers sharing covariates. It is not an observed individual treatment effect.

Identification requires consistency, conditional exchangeability, and positivity:

$$Y = AY(1) + (1-A)Y(0), \tag{2}$$
$$\{Y(1), Y(0)\} \perp\!\!\!\perp A \mid X, \tag{3}$$
$$0 < \Pr(A = 1 \mid X = x) < 1. \tag{4}$$

Consistency additionally requires a sufficiently well-defined intervention and an appropriate treatment version. The propensity score provides a way to represent observed treatment selection [12]. Define

$$e_0(x) = \Pr(A = 1 \mid X = x), \tag{5}$$
$$\mu_a(x) = \mathbb{E}(Y \mid A = a, X = x), \quad a \in \{0,1\}. \tag{6}$$

Under Eqs. (2)–(4), $\tau(x) = \mu_1(x) - \mu_0(x)$. Without exchangeability, the same observed contrast remains a covariate-adjusted association and does not identify an intervention effect.

In the application, A denotes technical-support subscription and $Y = 1$ corresponds to `Churn=No`. Technical support is a service-state proxy, not a recorded retention campaign. Contract and service variables lack established pre-treatment dates. Accordingly, empirical estimates are exploratory observational contrasts evaluated through a causal-estimation workflow. Table 1 specifies which fields enter that workflow and which claims their availability cannot support.

3.2 Cross-fitted doubly robust scores

For nuisance estimates $\widehat{e}(x)$ and $\widehat{\mu}_a(x)$, define the clipped propensity

$$\widehat{e}_\epsilon(x) = \min\{1 - \epsilon, \max(\epsilon, \widehat{e}(x))\}. \tag{7}$$

The augmented inverse-probability-weighted score is

$$\begin{aligned}\widehat{\phi}_i = {} & \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i) \\ & + \frac{A_i\{Y_i - \widehat{\mu}_1(X_i)\}}{\widehat{e}_\epsilon(X_i)} \\ & - \frac{(1-A_i)\{Y_i - \widehat{\mu}_0(X_i)\}}{1 - \widehat{e}_\epsilon(X_i)}.\end{aligned} \tag{8}$$

Both residual terms are essential. A difference between fitted outcome probabilities alone defines an outcome-regression contrast, not this doubly robust score [13].

Training records are divided into five folds. Each record receives nuisance predictions from models fitted on the other four folds. These out-of-fold predictions produce the training scores in Eq. (8). Full-training nuisance models subsequently produce scores on the held-out evaluation sample. Cross-fitting reduces direct reuse of a record for nuisance fitting and score construction [14]. All hyperparameters and the primary targeting fraction are fixed before held-out evaluation.

Table 1 Observed fields, analytical roles, and limits of identification

Role	Fields or definition	Interpretation and restriction
Treatment proxy	`TechSupport`: Yes versus No	Defined only among internet-service customers; no campaign assignment date or randomized intervention is recorded
Retention outcome	$Y = \mathbf{1}(\texttt{Churn=No})$	An observed binary status; no prospective outcome window can be reconstructed from individual records
Demographics	Gender, SeniorCitizen, Partner, Dependents	Enter the main specification; temporal precedence is plausible but not independently documented
Service covariates	PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, StreamingTV, StreamingMovies	Describe contemporaneous service configuration; their pre-treatment status is not established
Account covariates	Contract, PaperlessBilling, PaymentMethod	Describe current account arrangements; potential confounding and post-treatment conditioning remain concerns
Excluded billing fields	MonthlyCharges, TotalCharges	Charges may encode subscribed services; MonthlyCharges enters a separate sensitivity specification only
Descriptive time field	Tenure in months	Used for empirical distributions and cross-sectional retention proportions; does not supply an intervention-origin survival time
Excluded identifier	CustomerID	Used to check record uniqueness, never supplied to a model
Unavailable inputs	Intervention dates, censoring history, campaign costs, customer value	Required for prospective survival effects and realized financial evaluation

The following identity clarifies the source and boundary of robustness. It applies to nuisance functions evaluated on an independent observation under the identification assumptions.

Proposition 1 (Conditional score bias) *For candidate outcome functions m_a and a propensity function $\bar{e} \in (0,1)$, let $\phi(O; m, \bar{e})$ have the form of Eq. (8). Then*

$$\begin{aligned} &\mathbb{E}\{\phi(O; m, \bar{e}) \mid X = x\} - \tau(x) \\ &= \{\bar{e}(x) - e_0(x)\}\, R(x), \\ &R(x) = \frac{m_1(x) - \mu_1(x)}{\bar{e}(x)} + \frac{m_0(x) - \mu_0(x)}{1 - \bar{e}(x)}. \end{aligned} \tag{9}$$

Proof Conditioning on $X = x$ gives $\mathbb{E}[A\{Y - m_1(x)\} \mid X = x] = e_0(x)\{\mu_1(x) - m_1(x)\}$. The control residual has expectation $(1 - e_0(x))\{\mu_0(x) - m_0(x)\}$. Substituting these expectations and collecting the two regression errors yields Eq. (9). □

The identity vanishes when the propensity is correct or both outcome functions are correct. It does not generally vanish when both nuisance specifications are wrong [15]. Active clipping changes the candidate propensity and can introduce bias if the outcome functions are also misspecified. We therefore vary ϵ and report when clipping affects observations. In the main empirical specification, no fitted propensity falls outside $[0.025, 0.975]$.

3.3 Interpretable uplift ranking

Let $z(X)$ contain an intercept, reference-coded categorical indicators, and standardized numerical covariates. Separate penalized logistic regressions estimate e_0, μ_0, and μ_1. For each logistic model, parameters minimize

$$\begin{aligned} \mathcal{L}(\beta) = &\sum_i \{-y_i \log p_i - (1 - y_i) \log(1 - p_i)\} \\ &+ \frac{\lambda_L}{2} \sum_{j>0} \beta_j^2, \qquad p_i = \operatorname{expit}\{z(X_i)^\top \beta\}. \end{aligned} \tag{10}$$

The intercept is unpenalized and $\lambda_L = 1$. Separate outcome models use records within the corresponding treatment group.

The doubly robust learner fits a ridge regression to the cross-fitted scores:

$$\widehat{b} = \arg\min_{b} \left[\sum_{i \in \mathcal{T}} \{\widehat{\phi}_i - z(X_i)^\top b\}^2 + \lambda_R \sum_{j>0} b_j^2 \right], \tag{11}$$

where $\mathcal{T}$ is the training sample and $\lambda_R = 10$. Its ranking score is $s_{\mathrm{DR}}(x) = z(x)^\top \widehat{b}$. Ridge regularization stabilizes correlated design columns [16]. A linear second stage yields a regularized projection of conditional scores; it need not recover every nonlinear treatment-effect pattern [7]. Scores are therefore not treated as calibrated individual response probabilities.

We compare this score with two alternatives:

$$s_{\mathrm{risk}}(x) = 1 - \widehat{r}(x), \tag{12}$$

$$s_{\mathrm{T}}(x) = \widehat{\mu}_1(x) - \widehat{\mu}_0(x). \tag{13}$$

Here, $\widehat{r}$ is a pooled logistic retention model that excludes the treatment proxy. The T-learner shares the full-training outcome models used in score evaluation. The comparison therefore evaluates three practical rankings within a common, transparent nuisance specification. It does not constitute a search over all available machine-learning architectures.

Figure 1 connects training, diagnostics, held-out evaluation, and conditional extensions. The diagram distinguishes computations supported by the public records from prospective business and survival evaluation requiring further data.

3.4 Coverage, gain, and uncertainty

For evaluation size n_E and coverage q, let $k_q = \mathrm{round}(qn_E)$. A stable descending sort defines $\widehat{\pi}_s^q(X_i)$, selecting exactly k_q records. Stable ties follow the fixed evaluation-record order. We define the cumulative incremental-retention estimate

$$\widehat{G}_s(q) = \frac{1}{n_E} \sum_{i \in \mathcal{E}} \widehat{\pi}_s^q(X_i) \widehat{\phi}_i, \tag{14}$$

and the selected-group retention difference

$$\widehat{H}_s(q) = \frac{\sum_{i \in \mathcal{E}} \widehat{\pi}_s^q(X_i) \widehat{\phi}_i}{k_q}, \qquad k_q > 0. \tag{15}$$

The first measures incremental retention per member of the entire evaluation cohort. The second averages over selected customers. Multiplying either quantity by 100 expresses it in percentage-point units. Their relationship is $\widehat{G}_s(q) = (k_q/n_E)\widehat{H}_s(q)$. Distinguishing these denominators prevents a selected-group contrast from being interpreted as a gain for the whole population.

For descriptive ranking comparison, we compute

$$\widehat{Q}_s = \int_0^1 \{\widehat{G}_s(q) - q\widehat{\theta}\}\, dq, \quad \widehat{\theta} = \frac{1}{n_E} \sum_{i \in \mathcal{E}} \widehat{\phi}_i. \tag{16}$$

The integral uses a 51-point coverage grid and trapezoidal integration. This explicitly defined, doubly robust gain-area statistic is termed the DR gain area above random targeting here. It is reported on the probability scale and is not divided by an oracle or maximum attainable area. The random-targeting reference is $q\widehat{\theta}$; no additional random campaign is observed.

We resample evaluation rows 1,000 times using a nonparametric bootstrap [17]. Fitted models, scores, and selected-customer indicators remain fixed. For selected-group statistics, each replicate uses its resampled selected count as the denominator. The 2.5th and 97.5th percentiles form conditional intervals. Gain-curve bands are pointwise and are not simultaneous confidence bands. Paired ranking comparisons use the same resampled rows for both rules. These intervals quantify evaluation-sample variability while excluding training, model-selection, and unmeasured-confounding uncertainty.

3.5 Resource-constrained decisions

Suppose an eligible customer's retained value is v_i and intervention cost is c_i. The incremental net-value criterion is $v_i\tau(X_i) - c_i$. With at most K

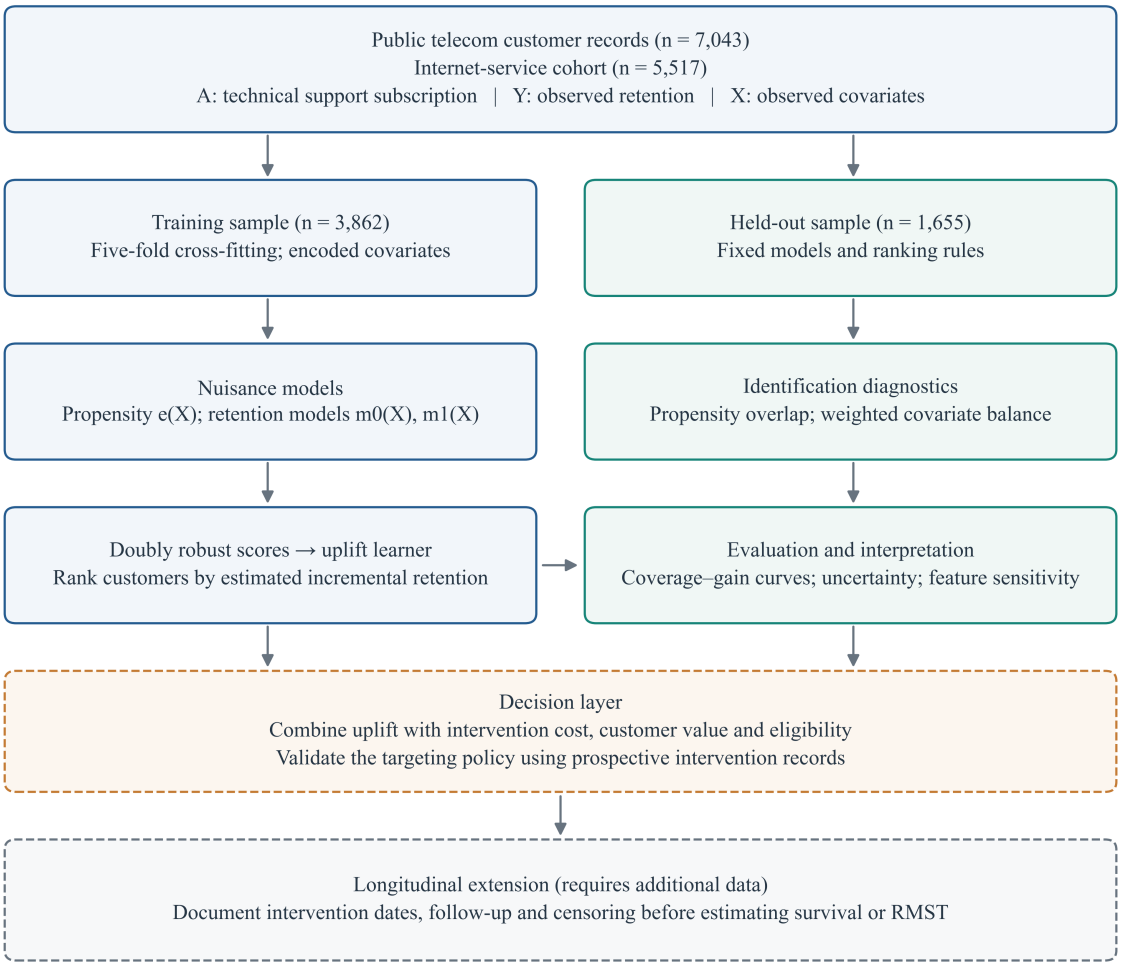


Fig. 1 Framework for evaluating customer retention targeting with cross-fitted doubly robust scores and interpretable uplift ranking. The public-data workflow uses technical-support subscription as a treatment proxy. Dashed boxes identify decision validation and longitudinal extensions requiring additional inputs

interventions, a plug-in policy solves

$$\widehat{\pi} \in \arg\max_{\pi \in \Pi_K} \frac{1}{n} \sum_{i=1}^{n} \pi_i \{ v_i \widehat{\tau}(X_i) - c_i \}, \qquad \Pi_K = \{ \pi \in \{0,1\}^n : \sum_i \pi_i \le K \}. \tag{17}$$

Under a cardinality constraint, the policy selects up to K positive net-value scores. Under a monetary budget, the constraint becomes $\sum_i c_i \pi_i \le B$, which is a different allocation problem. Neither optimization assumes that costs or values can be inferred from the benchmark's monthly charges.

A simple bound connects effect estimation to the decision objective. Let $0 \le v_i \le V_{\max}$, let $q = K/n$, and define $\Gamma(\pi) = n^{-1} \sum_i \pi_i \{ v_i \tau_i - c_i \}$. For an oracle policy π^* and the plug-in maximizer in Eq. (17),

$$\Gamma(\pi^*) - \Gamma(\widehat{\pi}) \le 2 V_{\max} \sqrt{q} \left\{ \frac{1}{n} \sum_i (\widehat{\tau}_i - \tau_i)^2 \right\}^{1/2}. \tag{18}$$

This finite-sample accounting bound follows from the optimality of $\widehat{\pi}$ and the Cauchy–Schwarz inequality. Its derivation appears in Appendix A. It is an implication of the decision formulation, rather

than a claim of a new general policy-learning rate. It shows why effect-estimation error matters even when the final policy is simple [8].

For constant customer value v and constant cost ratio $\kappa = c/v$, the value-normalized estimate is $\widehat{G}_s(q) - (k_q/n_E)\kappa$. Thus, the point-estimate break-even ratio is $\widehat{H}_s(q)$. A conservative conditional check instead compares κ with the lower interval endpoint. Costs and values are absent from the public data, so we report this relationship without assigning invented monetary amounts.

3.6 Model interpretation and longitudinal boundaries

For feature group j, we permute its encoded columns jointly across held-out customers. Grouping preserves a categorical variable's internal coding. Across 30 permutations, the sensitivity measure is

$$I_j = \frac{1}{30 n_E} \sum_{b=1}^{30} \sum_{i \in \mathcal{E}} \left| s_{\mathrm{DR}}(X_i^{(b,j)}) - s_{\mathrm{DR}}(X_i) \right|. \tag{19}$$

This is prediction sensitivity, not a loss-based importance measure or a SHAP decomposition. Correlated features and unrealistic shuffled combinations can affect it [10]. Contract-group summaries use averages of held-out doubly robust scores with conditional bootstrap intervals. They describe estimated group contrasts without identifying each customer's counterfactual response type.

The labels "persuadable", "always retained", and "unavoidable churner" concern the joint potential outcomes of an individual. Observational records reveal only one potential outcome, and separate marginal response models do not identify that joint distribution. We therefore avoid assigning definitive counterfactual customer types. Positive or negative uplift scores instead indicate model-based ranking priorities requiring validation.

For a genuine longitudinal extension, let $T(a)$ denote churn time from a documented intervention origin, and define $S_a(t) = \Pr\{T(a) > t\}$. The restricted mean survival-time contrast at horizon t_0 is

$$\Delta_{\mathrm{RMST}}(t_0) = \int_0^{t_0} \{S_1(t) - S_0(t)\}\, dt. \tag{20}$$

This quantity compares expected retained duration over a specified horizon [18]. Estimation additionally requires intervention timing, an at-risk population, event information, and an appropriate censoring model. Current accumulated tenure does not supply these elements. Consequently, no empirical survival-treatment effect or RMST improvement is reported in this study.

4 Data and evaluation design

4.1 Public benchmark and analysis cohort

The analysis uses the 21-column IBM Telco Customer Churn CSV distributed through IBM's public repository [19]. IBM describes its telco sample as a fictional-company data resource [20]. We use the CSV's observed churn status and do not assign calendar years or reconstruct monthly operational trends. The resource illustrates a computational workflow rather than representing a prospective intervention study.

The original file contains 7,043 unique customer records. We exclude 1,526 records with `InternetService=No`, because technical-support subscription is undefined for this service configuration. The resulting cohort contains 5,517 customers, including 2,044 technical-support subscribers and 3,473 non-subscribers. No further records are deleted. Missing or blank total charges do not require imputation because TotalCharges is excluded from all fitted models.

Fifteen observed covariates enter the main analysis, as specified in Table 1. Tenure, MonthlyCharges, TotalCharges, CustomerID, treatment, and outcome are excluded from the covariate design. The exclusions prevent direct reuse of outcomes, avoid interpreting tenure as a baseline time origin, and limit direct billing-based encoding of the subscribed service. They do not establish that every remaining field is pre-treatment.

4.2 Partition, preprocessing, and numerical implementation

We reserve approximately 30% of each treatment–outcome stratum for held-out evaluation using seed 20261009. Rounding within strata gives 3,862 training and 1,655 evaluation records, as shown in Table 2. The training set is further partitioned into

Table 2 Analysis cohort and fixed training–evaluation partition

Characteristic	Entire cohort	Training	Held-out
Customers, n	5,517	3,862	1,655
Technical support, n (%)	2,044 (37.05)	1,431 (37.05)	613 (37.04)
Retained, n (%)	3,761 (68.17)	2,633 (68.18)	1,128 (68.16)
Senior status, n (%)	1,090 (19.76)	758 (19.63)	332 (20.06)
Tenure, median [IQR]	30 [9,56]	30 [9,56]	30 [9,56]

Counts are customer records. Tenure is reported as median [interquartile range] in months and is used only descriptively. The partition is stratified by the joint treatment–retention label

five folds with the same joint-label stratification. No evaluation outcome is used to fit a nuisance model or ranking model.

Categorical variables use one-hot reference coding. The category vocabulary is taken from the observed cohort; no category frequencies or outcomes enter this mapping. SeniorCitizen is standardized using training-set statistics, and the resulting transformation is reused in every fold and on the evaluation set. The main design contains 20 columns including the intercept. Logistic parameters are fitted by penalized Newton updates, with an iteration limit of 100 and a maximum parameter-update tolerance of 10^{-7}. Ridge coefficients solve the penalized normal equations directly.

The pooled logistic retention classifier provides a predictive reference, with retention as the positive class and a 0.5 classification threshold. We report its area under the receiver operating characteristic curve and F1 score separately from targeting metrics. An uplift score has a different target and is not assigned an artificial classification F1 score. Software calculations use Python, NumPy 2.5.3, and pandas 3.0.1. Figures are generated using Matplotlib 3.11.2 [21], with numerical outputs checked against the tables.

4.3 Overlap, balance, and sensitivity checks

Propensity overlap is examined using empirical distributions for the two support groups. For a numerical covariate or categorical indicator Z, balance is summarized by

$$\mathrm{SMD}(Z) = \frac{|\overline{Z}_1 - \overline{Z}_0|}{\{(s_1^2 + s_0^2)/2\}^{1/2}}. \tag{21}$$

The weighted version replaces group means by inverse-probability-weighted means and retains the original pooled standard deviation. Categorical variables report the maximum across all levels, including the reference level. Standardized differences diagnose measured-covariate imbalance rather than testing causal identification [22]. We display 0.1 as a descriptive reference, not a universal acceptance threshold.

For the diagnostic cohort, training records receive out-of-fold propensity predictions and evaluation records receive full-training predictions. We use weights $A_i/\widehat{e}_\epsilon(X_i) + (1 - A_i)/\{1 - \widehat{e}_\epsilon(X_i)\}$. Group-specific effective sample size is $(\sum w_i)^2/\sum w_i^2$. No propensity trimming or overlap-based record deletion is applied. Figure 2 reports the resulting diagnostics.

Two sensitivity analyses probe modelling choices. We vary the clipping threshold over $\{0.01, 0.025, 0.05, 0.10\}$, reconstruct training scores, and refit the ridge learner. Separately, we compare the main covariates with a demographics-only specification and a specification adding standardized MonthlyCharges. Each alternative repeats nuisance fitting and score learning using the same partition. The demographics-only analysis may omit confounders; the billing analysis may condition on treatment-related information. Neither alternative is designated a preferred causal adjustment set.

4.4 Semi-synthetic robustness experiment

The public treatment proxy does not provide observed causal ground truth. We therefore conduct a separate experiment with generated binary treatment and retention labels. Each repetition samples 3,000 covariate rows with replacement from the analysis cohort. The first 2,000 rows form the training sample and the remaining 1,000 form an independent evaluation sample. An eight-column design z contains the intercept, standardized senior status, fiber-optic service, online

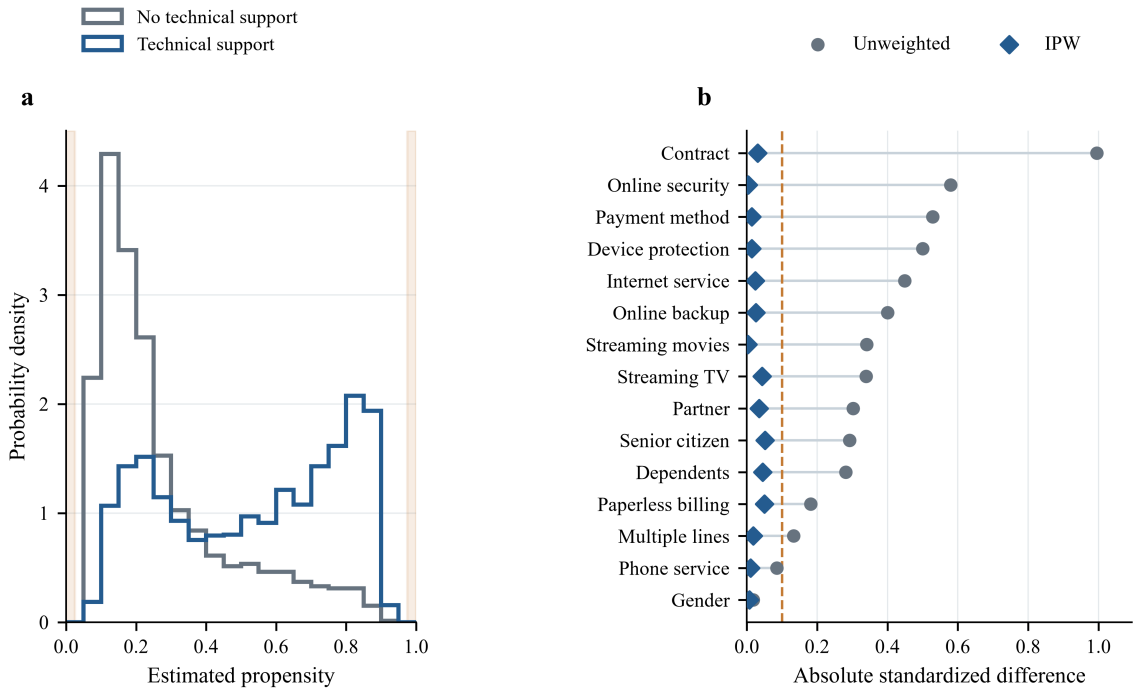


Fig. 2 Observed treatment-selection diagnostics in the internet-service cohort. (a) Propensity distributions by technical-support subscription; shaded boundary regions correspond to the primary clipping range. (b) Absolute standardized differences before and after inverse-probability weighting, with categorical variables summarized across all levels. The dashed line marks 0.1; balance does not validate pre-treatment timing or eliminate unmeasured confounding

security, one-year and two-year contract indicators, paperless billing, and partner status.

Treatment and potential-outcome means follow

$$e_0(z) = \text{expit}(z^\top \beta_e), \tag{22}$$

$$\mu_0(z) = \text{expit}(z^\top \beta_0), \tag{23}$$

$$\mu_1(z) = \text{expit}\{z^\top (\beta_0 + \beta_\Delta)\}. \tag{24}$$

Appendix B specifies every coefficient. We draw $A \sim \text{Bernoulli}(e_0)$ and $Y \mid A, z \sim \text{Bernoulli}(\mu_A)$ independently across sampled rows. All treatment and outcome labels in this experiment are generated; the original support and churn labels are not used.

Correctly specified nuisance models use all eight columns. Misspecified models use an intercept only, omitting the covariates driving treatment and retention. We cross these choices to produce four regimes and compare outcome regression, inverse-probability weighting, and the doubly robust score. Nuisance models are fitted on the independent training sample, so out-of-fold predictions are unnecessary for this average-effect comparison. The ground-truth target in each repetition is the evaluation sample's average $\mu_1(z) - \mu_0(z)$.

We execute 200 repetitions using a dedicated generator with seed 20261011. For each estimator, we report average error in percentage points and root mean squared error (RMSE). Monte Carlo standard errors quantify uncertainty in estimated average bias across repetitions. The experiment examines a specified finite-sample mechanism; it is not an external validation of the technical-support proxy or a demonstration of universal learner superiority.

4.5 Partition stability and decision validation

We repeat the joint-label-stratified 70/30 partition 30 times using a dedicated generator with seed

20261012. Each repetition standardizes senior status using its training rows, reconstructs five training folds, and refits every model with unchanged penalties. We summarize top-20% contrasts by median, interquartile range, and range. These are sensitivity summaries over overlapping partitions of the same cohort, not confidence intervals or independent population replications.

A separate experiment evaluates conditional effects and decisions under three specified mechanisms. Each mechanism uses 200 repetitions with 2,000 training and 1,000 evaluation rows, and a dedicated generator with seed 20261013. The logistic mechanism follows Eqs. (22)–(24). Let F, M, P, H, S, O denote fiber-optic service, monthly contract, paperless billing, partner status, standardized senior status, and online security. The interaction mechanism adds $0.8FS - 0.65OP$ to the baseline logit and $1.2FM - 0.9PH + 0.5SO$ to the treatment logit difference. The weak-overlap mechanism multiplies the propensity logit by three while retaining the original outcome logits.

Outcome and propensity models use the eight original design columns; in the interaction mechanism they also include (FM, PH, SO, FS, OP). Within each repetition, both DR rankings share the same five-fold nuisance scores. DR linear uses the original eight columns in its ridge stage; DR interactions additionally uses those five products. This isolates second-stage specification within a mechanism. The T-learner uses the same full-training outcome models. The churn-risk ranking uses a pooled retention model with the same nuisance design. No hyperparameter search is performed.

Known generating means give $\tau_i = \mu_1(X_i) - \mu_0(X_i)$. We report CATE RMSE for effect learners and the true average effect among the top 20% for all rankings. For normalized cost κ and capacity $K = 200$, define

$$\Gamma_\kappa(\pi) = \frac{1}{n_E} \sum_{i \in \mathcal{E}} \pi_i(\tau_i - \kappa), \qquad D_\kappa = 100\{\Gamma_\kappa(\pi^*) - \Gamma_\kappa(\widehat{\pi})\}. \tag{25}$$

The oracle selects up to K positive true net-effect scores. Effect learners select up to K predicted effects exceeding κ; the churn-risk baseline fills the same capacity using risk alone. All are feasible under Eq. (17). We evaluate $\kappa \in \{0, 0.05, 0.10, 0.20, 0.30\}$ and report a common reference at 0.05. These ratios are simulation inputs, not recorded business costs. D_κ measures lost value per 100 evaluation customers in units of retained-customer value. Monte Carlo standard errors describe variability of the reported simulation means. Original treatment and churn labels are never used as outcomes in these simulations.

4.6 Assistance and reproducibility

ChatGPT (OpenAI) assisted with manuscript restructuring, English drafting, analysis-code implementation, and LaTeX preparation. All reported numerical results were generated by executing the stated calculations; generated prose was not treated as numerical evidence. The source includes executable analysis and simulation code. Scientific interpretation, factual verification, and submission accountability remain the responsibility of the author.

5 Results

5.1 Selection differences and measured balance

The two support groups differed substantially before adjustment. The largest unweighted absolute standardized difference was 0.995, whereas the largest weighted difference was 0.052. All reported weighted differences were below the descriptive 0.1 reference in Fig. 2. Effective sample sizes were 1,291.2 for support subscribers and 2,471.4 for non-subscribers, compared with original group sizes of 2,044 and 3,473. These reductions indicate that weighting changes the information contributed by the observed records.

The unadjusted support-group retention difference was 26.47 percentage points across the cohort. On the held-out sample, the average doubly robust score was 10.93 percentage points, with a conditional interval of $[5.65, 16.09]$. These quantities concern different adjustment procedures and samples, so their difference is not interpreted as an exact estimate of confounding bias. The positive adjusted contrast also does not resolve the unknown temporal ordering of support, contract, and service variables.

5.2 Held-out targeting comparisons

The pooled logistic retention model achieved an AUC of 0.7882 and an F1 score of 0.8080. These predictive metrics characterize its ability to reproduce observed retention labels. Targeting evaluation used a separate doubly robust score and selected 331 customers for each top-20% ranking. Table 3 reports the targeting estimates, and Fig. 3 displays their coverage profiles and uncertainty.

The churn-risk, T-learner, and DR rankings produced top-20% point estimates of 18.18, 13.65, and 12.52 percentage points, respectively. Their DR gain area above random targetings were 0.02175, 0.00939, and 0.00636. Thus, the DR learner did not have the largest point estimate under this specification. Its top-20% interval, $[-5.43, 28.75]$, included zero and covered a broad range of possible evaluation-sample contrasts.

The paired DR-minus-risk difference was -5.67 percentage points, with interval $[-17.36, 6.12]$. The paired DR-minus-T-learner difference was -1.13, with interval $[-9.71, 8.39]$. These comparisons did not establish a ranking advantage for the DR learner. They also illustrate why separate intervals for each method should not substitute for an explicitly paired comparison.

The distinction between selected-group and whole-cohort gains has a practical consequence. The DR top-20% contrast of 12.52 corresponds to approximately 2.50 incremental retained customers per 100 evaluation customers under the identifying interpretation. It does not correspond to a 12.52-point increase across the whole cohort. Under constant value and cost, the point-estimate break-even cost-to-value ratio would be 0.1252. Its negative conditional lower endpoint does not support a positive spending threshold at the reported uncertainty level.

5.3 Tenure patterns and interpretation

Support subscribers had a different observed tenure distribution from non-subscribers, as shown in Fig. 4a. The plot reports an empirical cumulative distribution of accumulated tenure. It contains neither a documented intervention origin nor treatment-specific follow-up risk sets. It therefore cannot be read as a Kaplan–Meier estimate of a support intervention's survival benefit.

Within six tenure intervals, observed retention proportions generally increased with accumulated tenure. Figure 4b reports these proportions with 95% Wilson intervals [23]. The summaries are unadjusted and cross-sectional. Selection into longer tenure, differences in customer composition, and the timing of service adoption can all contribute to these patterns. No event-time causal effect is calculated from them.

Grouped permutation sensitivity was largest for payment method, online security, and senior status in the fitted DR score. Figure 5a shows the ten largest sensitivities. These values indicate how the model's predictions change after shuffling a feature group. They neither identify causal drivers of response nor establish that a feature is an appropriate intervention target.

Contract-group score summaries varied in both magnitude and precision. Monthly-contract customers had an estimated contrast of 17.11 percentage points, with interval $[9.15, 25.45]$ and $n = 1,008$. The corresponding estimates were -0.34 for one-year contracts ($n = 335$) and 3.08 for two-year contracts ($n = 312$). Their intervals, $[-8.25, 7.98]$ and $[-2.85, 9.53]$, included zero. Figure 5b reports these estimates without assigning individual counterfactual customer categories.

5.4 Sensitivity to clipping and covariate specification

Clipping thresholds of 0.01, 0.025, and 0.05 did not change any cohort propensity prediction. Their DR top-20% point estimate remained 12.52 percentage points. At a threshold of 0.10, 426 cohort predictions were clipped and the refitted DR ranking produced an estimate of 13.77. Figure 6a displays these comparisons. Their conditional intervals remained broad, so the higher point estimate does not establish an improved targeting policy.

Covariate choice had a larger effect, as shown in Table 4. The demographics-only specification produced an average score of 23.32 percentage points, compared with 10.93 under the main adjustment set. Adding monthly charges produced an average score of 8.87 and moved held-out propensity predictions as low as 0.0062 and as high as 0.9949. Ninety-eight held-out propensities then required clipping at 0.025.

The billing specification increased the DR top-20% estimate to 17.69 and narrowed its conditional

Table 3 Held-out ranking evaluation using a common doubly robust score

Ranking rule	Selected n	Top-20% contrast	Conditional 95% interval	DR gain area
Churn-risk ranking	331	18.18	[0.67, 35.19]	0.02175
T-learner uplift	331	13.65	[−3.34, 28.32]	0.00939
DR-learner uplift	331	12.52	[−5.43, 28.75]	0.00636

Contrasts and interval endpoints are percentage points among the selected customers. DR gain areas follow Eq. (16) and are reported on the probability scale. Intervals use 1,000 evaluation-row bootstrap samples with fitted models and selected-customer indicators fixed; they exclude training uncertainty

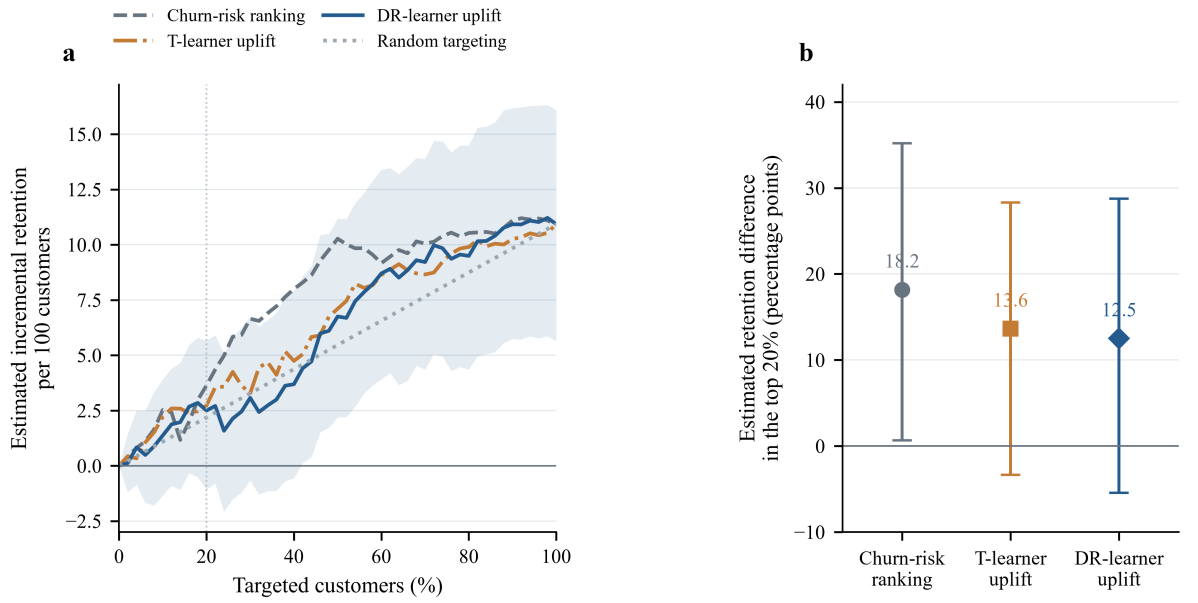


Fig. 3 Exploratory targeting evaluation on 1,655 held-out customers. (a) Estimated cumulative incremental retention per 100 customers across coverage; the reference represents random targeting under the same estimated average contrast. Shading gives pointwise conditional 95% bootstrap intervals for the DR ranking. (b) Selected-group retention differences for the top 20%, with conditional 95% intervals. Fitted models and targeting indicators remain fixed in all bootstrap samples

interval to [11.61, 23.47]. However, improved apparent precision did not resolve the status of monthly charges as a possible descendant or direct encoding of subscribed services. This sensitivity demonstrates that more favourable targeting numbers can accompany a less defensible covariate interpretation. The main specification consequently excludes both billing fields, while acknowledging that remaining service variables also lack verified temporal precedence.

5.5 Stability across training partitions

Table 5 reports the 30 complete refits. The DR median was 13.40 percentage points, with interquartile range [9.68, 20.83] and range [2.06, 27.48]. The variation complements the fixed-model conditional intervals: it exposes dependence on the training partition but does not provide a population confidence interval. Repartitioning does not resolve the uncertain intervention definition or establish external validity.

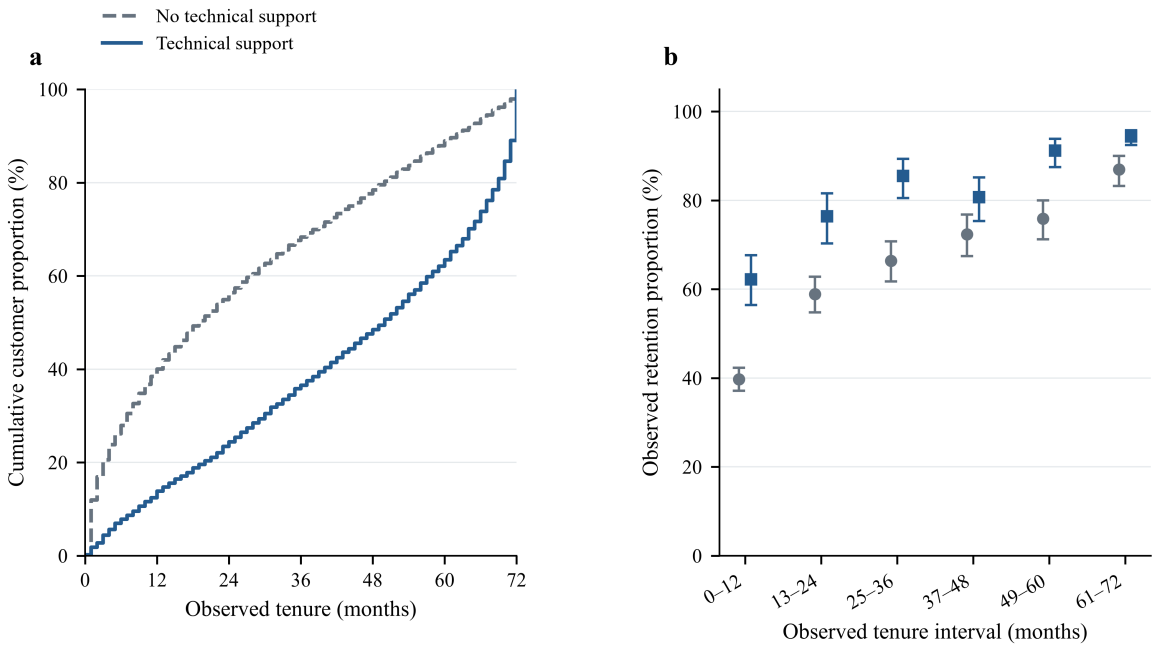


Fig. 4 Descriptive tenure and retention patterns by technical-support subscription. (a) Empirical cumulative distributions of observed tenure. (b) Observed retention proportions within tenure intervals, with 95% Wilson intervals. These cross-sectional summaries do not estimate post-intervention survival curves or restricted mean survival-time effects

Table 4 Sensitivity of the DR ranking to observed covariate specification

Specification	Mean score	Top-20% contrast [95% interval]	Held-out propensity range	Clipped n
Main covariates	10.93	12.52 [−5.43, 28.75]	[0.0607, 0.9045]	0
Demographics only	23.32	19.63 [8.75, 30.22]	[0.1871, 0.5094]	0
Main plus monthly charges	8.87	17.69 [11.61, 23.47]	[0.0062, 0.9949]	98

Score contrasts are percentage points. Each specification repeats five-fold nuisance fitting and ridge score learning with the same split. The clipping count concerns 1,655 held-out records at $\epsilon = 0.025$. The demographics-only model omits potentially relevant service covariates; monthly charges may encode the subscribed service. Neither alternative establishes a valid causal adjustment set

5.6 Robustness under known semi-synthetic ground truth

The simulation results separated nuisance robustness from the unresolved identification problems of the public proxy. Table 6 reports estimator bias, Monte Carlo standard error, and RMSE for all four specification regimes. Figure 7 displays the same comparisons. The ground-truth average effect was computed directly from the generating probabilities within each evaluation sample.

When both nuisance specifications were correct, outcome regression had RMSE 2.13 percentage points, compared with 3.19 for the doubly robust estimator. Thus, the augmented estimator did not universally minimize finite-sample error. The correctly specified parametric outcome model was particularly effective under this generating mechanism.

When only the propensity specification was correct, outcome-regression bias was 4.85 percentage points, whereas doubly robust bias was -0.35. When only the outcome specification was correct, IPW bias was 4.16, compared with -0.23 for the doubly robust estimator. These results matched the qualitative cancellation mechanism in Proposition 1, while retaining finite-sample estimation error.

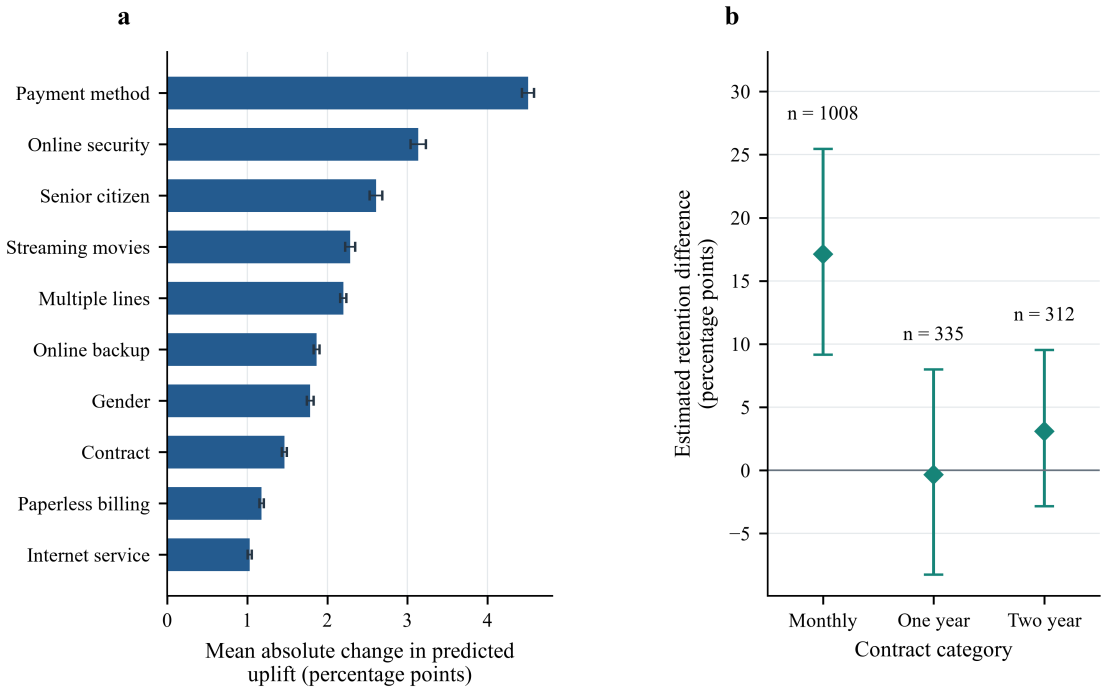


Fig. 5 Model interpretation and exploratory contract-group contrasts. (a) Mean absolute changes in predicted uplift after grouped feature permutation, expressed in percentage points; error bars give standard deviations across 30 permutations. (b) Mean held-out doubly robust scores by contract category, with conditional 95% bootstrap intervals and subgroup counts. Neither feature sensitivity nor subgroup membership identifies an individual causal response type

Table 5 Top-20% contrasts across 30 repeated partitions

Ranking	Median	Interquartile range	Minimum–maximum
Churn-risk ranking	18.96	[15.42, 23.90]	[3.68, 31.82]
T-learner uplift	16.25	[11.56, 20.76]	[4.90, 29.70]
DR-learner uplift	13.40	[9.68, 20.83]	[2.06, 27.48]

All values are percentage points. Each partition repeats preprocessing, five-fold nuisance fitting, and ranking-model fitting on the same cohort. Ranges describe partition sensitivity and are not confidence intervals

When both specifications omitted the generating covariates, doubly robust bias increased to 4.37 percentage points and its RMSE reached 5.40. This failure is consistent with the nonzero product of nuisance errors in Eq. (9). The experiment therefore supports a conditional robustness claim, rather than robustness to arbitrary omitted-variable bias.

5.7 Conditional-effect error and decision loss

Table 7 and Fig. 8 separate effect recovery from decision quality across the 600 additional generated-data experiments. In the interaction mechanism, expanding the DR second stage reduced mean CATE RMSE from 11.32 to 7.96 percentage points and reference decision loss from 0.43 to 0.20 per 100 customers. Under weak overlap, the same expansion increased RMSE from 9.34 to 11.66 and decision loss from 0.63 to 0.80. On average, 5.53% of weak-overlap evaluation rows had true propensities outside the main clipping interval. The T-learner had the lowest reference decision loss in all three mechanisms: 0.24, 0.16, and 0.30. At the reference cost ratio, all fitted policies used the full 20% capacity; the cost-grid analysis separately evaluates the possibility of withholding treatment as the threshold increases. These

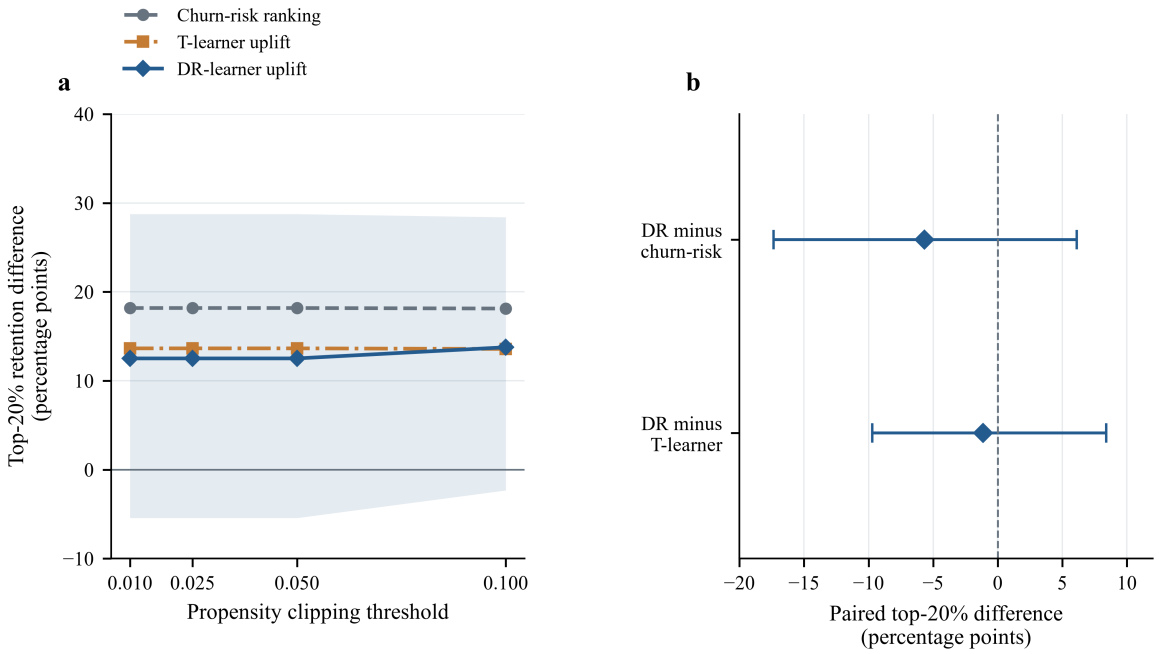


Fig. 6 Targeting sensitivity and paired comparisons. (a) Top-20% retention differences across propensity clipping thresholds, with the DR learner refitted after reconstructing training scores; shading gives conditional 95% intervals for the DR rule. (b) Paired DR-minus-baseline top-20% differences using common evaluation-row resamples. The reference line denotes zero difference. Intervals exclude uncertainty from model fitting and covariate-set selection

Table 6 Average-effect estimation under semi-synthetic nuisance misspecification

Nuisance specification	Estimator	Bias	MCSE of bias	RMSE
Both correct	Outcome regression	0.31	0.15	2.13
	IPW	-0.50	0.40	5.69
	Doubly robust	-0.41	0.22	3.19
e correct; m omitted	Outcome regression	4.85	0.16	5.37
	IPW	-0.50	0.40	5.69
	Doubly robust	-0.35	0.25	3.48
e omitted; m correct	Outcome regression	0.31	0.15	2.13
	IPW	4.16	0.38	6.82
	Doubly robust	-0.23	0.22	3.11
Both omitted	Outcome regression	4.85	0.16	5.37
	IPW	4.16	0.38	6.82
	Doubly robust	4.37	0.23	5.40

All entries are percentage points across 200 generated-data repetitions, each with 2,000 training and 1,000 evaluation rows. Correct models include all eight generating design columns. Omitted-covariate models contain only an intercept. The target is the true average effect within each generated evaluation sample; labels and outcomes are simulated, not observed business interventions

comparisons concern the stated mechanisms and shared nuisance designs. They do not identify a universally preferred learner.

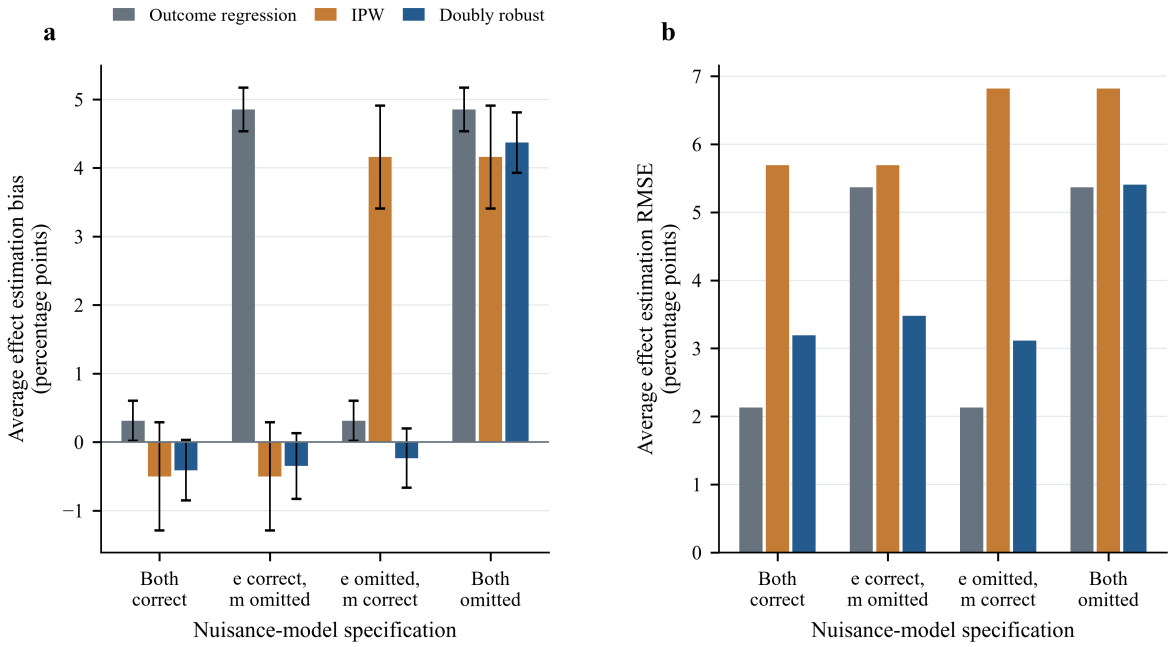


Fig. 7 Semi-synthetic average-effect estimation with known ground truth across 200 repetitions. (a) Mean estimation bias; error bars show 1.96 Monte Carlo standard errors of the mean bias. (b) Root mean squared estimation error. Correct nuisance specifications contain the generating covariates; omitted specifications are intercept-only. Treatment and retention labels are generated from the stated mechanism, independently of the benchmark's observed labels

Table 7 Conditional effects and capacity-constrained decisions with known ground truth

Mechanism	Ranking	CATE RMSE (pp)	True top-20% contrast (pp)	Decision loss per 100
Logistic mechanism	Churn-risk ranking	—	23.00 (0.10)	0.99 (0.02)
	T-learner uplift	5.47 (0.09)	26.72 (0.09)	0.24 (0.02)
	DR linear	5.97 (0.09)	26.25 (0.09)	0.34 (0.02)
	DR interactions	7.50 (0.10)	25.91 (0.11)	0.41 (0.02)
Interaction mechanism	Churn-risk ranking	—	20.92 (0.52)	5.83 (0.10)
	T-learner uplift	7.06 (0.11)	49.30 (0.08)	0.16 (0.01)
	DR linear	11.32 (0.06)	47.93 (0.13)	0.43 (0.03)
	DR interactions	7.96 (0.10)	49.06 (0.07)	0.20 (0.01)
Weak overlap	Churn-risk ranking	—	24.26 (0.07)	0.74 (0.01)
	T-learner uplift	6.26 (0.12)	26.43 (0.10)	0.30 (0.02)
	DR linear	9.34 (0.17)	24.81 (0.18)	0.63 (0.04)
	DR interactions	11.66 (0.19)	23.92 (0.22)	0.80 (0.04)

Entries are Monte Carlo means (standard errors) across 200 repetitions per mechanism, with $n_E = 1,000$, $K = 200$, and $\kappa = 0.05$. The dash indicates that a risk score is not a conditional-effect estimate. Decision loss is defined in Eq. (25); original churn and treatment outcomes are replaced by generated labels

6 Discussion

6.1 What the framework establishes

The framework makes the retention decision objective explicit and exposes where its supporting evidence is weak. The public-data application

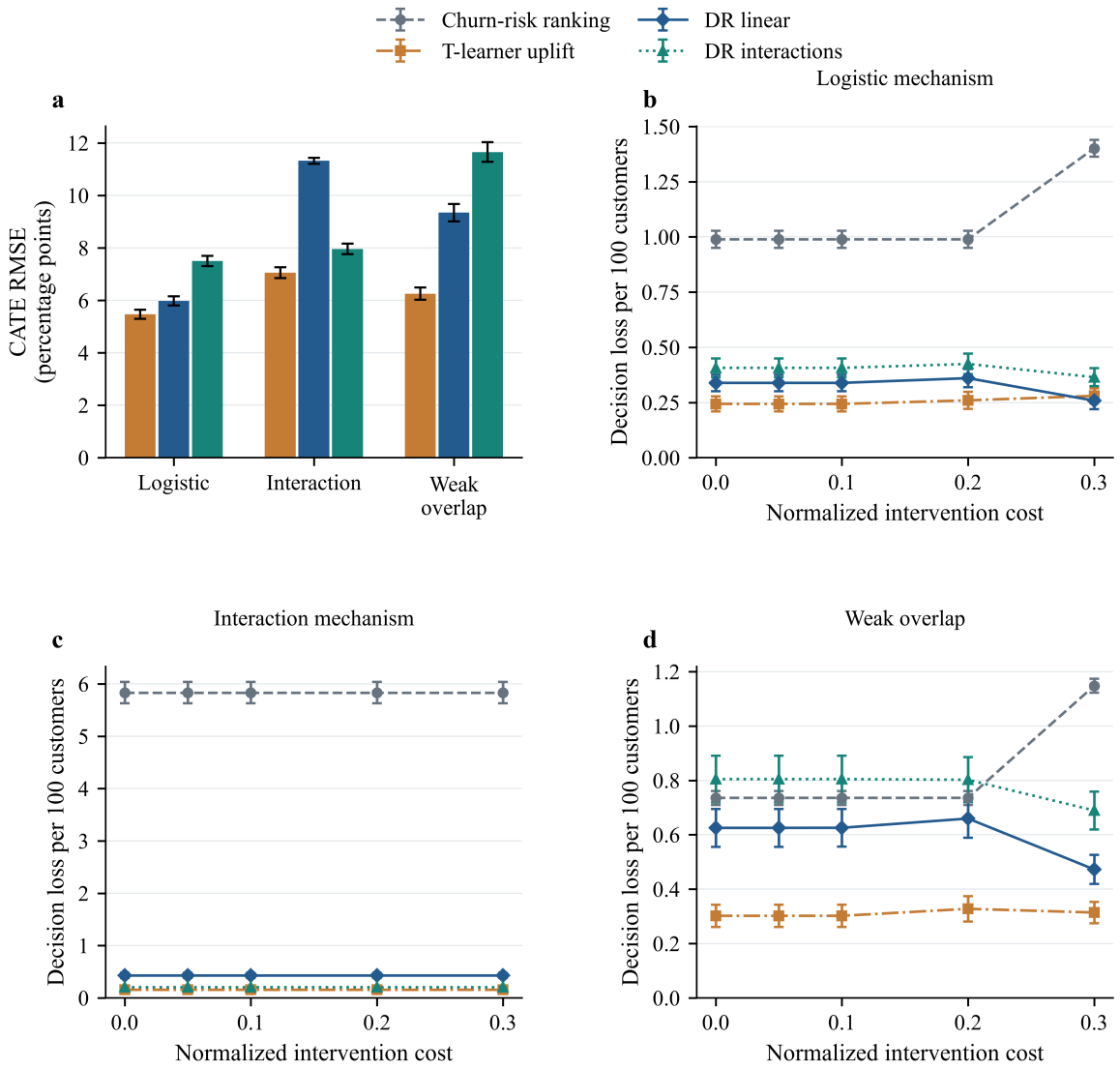


Fig. 8 Conditional-effect recovery and decision loss under specified simulation mechanisms. (a) Mean CATE RMSE; error bars show 1.96 Monte Carlo standard errors across 200 repetitions. (b–d) Oracle-relative decision loss over the illustrative normalized-cost grid under logistic, interaction, and weak-overlap mechanisms; error bars show 1.96 Monte Carlo standard errors. Capacity is at most 20%. The risk baseline fills capacity, whereas effect learners omit predicted effects below cost. Treatment and retention labels are generated, and these costs do not represent a recorded retention campaign

demonstrates a reproducible sequence from variable definitions to cross-fitted scores, diagnostics, targeting rules, and uncertainty summaries. The simulations verify the nuisance-robustness mechanism and examine how second-stage specification and overlap affect effect recovery and decision loss. Neither component establishes that technical-support subscription caused the observed retention differences.

The repeated partitions and known-effect experiments show why average-effect robustness should be assessed separately from ranking reliability. The controlled interaction comparison holds nuisance predictions fixed while changing the score-regression basis. Its improved effect recovery does not carry over to weak overlap, where residual weighting and added model dimensions increase error. This is an empirical specification trade-off

within the stated design, not a new robustness theorem.

The ranking results also limit claims about algorithmic performance. The DR learner did not outperform the two baselines in the observed point estimates, and paired intervals included zero. This finding does not invalidate the score's robustness identity. A low-dimensional second-stage model, residual-score variability, modest evaluation size, and the particular treatment proxy can all affect ranking quality. Robustness of nuisance adjustment and superiority of a learned ranking are separate properties.

The covariate sensitivity results are especially relevant to applied data science. Adding monthly charges made technical-support status easier to predict and changed the apparent precision of targeting gains. Because billing may incorporate the subscribed service, this improvement can reflect conditioning on information affected by treatment. Conversely, restricting adjustment to demographics can leave service-selection differences uncontrolled. Data availability therefore cannot replace a reasoned adjustment strategy with documented timing.

6.2 Implications for subscription decision support

An operational retention system should first define a manipulable action and a prospective outcome window. Pre-action covariates then support an effect estimate, while customer value, intervention cost, and eligibility determine the feasible policy. Equation (17) makes those commercial inputs explicit. Its value objective is not represented by classification accuracy, and its inputs should not be fabricated from a convenient billing field.

Uncertainty can change the action even when a score is positive. The DR top-20% lower interval endpoint was negative in this application. Under the constant-value illustration, that endpoint does not justify a positive cost-to-value threshold. A practical next step would be a randomized pilot with a documented treatment, baseline covariates, and common follow-up. The pilot should compare a pre-specified targeting rule with an appropriate allocation baseline and evaluate net value using recorded campaign costs.

Interpretability should similarly remain tied to the decision quantity. Feature sensitivity explains the fitted uplift score, whereas subgroup summaries describe estimated contrasts within observed customer categories. Neither can identify a customer's unobserved pair of potential outcomes. Treating such outputs as hypotheses for campaign design avoids converting model explanations into unsupported behavioural mechanisms.

6.3 Limitations and future evaluation

The study has five specific limitations. First, the public resource is a fictional-company sample and contains a service-state treatment proxy. External campaign effectiveness cannot be established from this resource. Second, service and contract covariates are contemporaneous, leaving exchangeability and temporal precedence unverified. Good measured balance does not remove these concerns.

Third, the primary comparison uses one fixed partition and low-dimensional penalized models, supplemented by 30 complete refits. The conditional bootstrap does not incorporate training variation, hyperparameter uncertainty, or the effects of searching over covariate specifications. The sensitivity analyses should therefore be interpreted descriptively, rather than as independent confirmatory experiments. Fourth, subgroup exploration and permutation sensitivity can be unstable under correlated covariates or smaller sample sizes. No multiplicity-adjusted confirmatory subgroup claim is made.

Fifth, the generated-data experiments cover four omission regimes and three specified decision mechanisms, but retain observed covariates and fixed parametric nuisance learners. They do not establish robustness to hidden confounding or transport to another market. The interaction features follow the known generating design, so their benefit does not demonstrate automatic recovery of unknown business mechanisms. Future evaluations require broader learners and genuinely independent intervention datasets. Longitudinal validation additionally requires event origin and censoring information before Eq. (20) can be estimated.

7 Conclusion

Customer retention targeting requires an evaluation of incremental response, uncertainty, and decision constraints. We present a framework that

combines cross-fitted doubly robust scores, interpretable ranking, observed-covariate diagnostics, and held-out policy comparison. Its public-data application shows favourable measured balance but does not establish a ranking advantage for the DR learner or a causal effect of support subscription. Sensitivity to billing variables further demonstrates why numerical improvements require scrutiny of variable timing.

Controlled semi-synthetic experiments show the expected robustness pattern when one nuisance specification is correct and substantial bias when both omit relevant covariates. Repeated partitions expose ranking variation, while known-effect decision experiments show that richer second-stage models can improve interaction recovery and increase error under weak overlap. The combined evidence supports evaluating identification, estimation, and decisions separately before adopting a targeting rule. Prospective deployment should proceed through documented interventions, baseline measurement, recorded costs, and independent outcome follow-up.

A Decision-bound derivation

Let $\widehat{\Gamma}(\pi) = n^{-1}\sum_i \pi_i\{v_i\widehat{\tau}_i - c_i\}$. The optimality of $\widehat{\pi}$ implies $\widehat{\Gamma}(\pi^*) - \widehat{\Gamma}(\widehat{\pi}) \le 0$. Adding and subtracting these two estimated values gives

$$\begin{aligned}\Gamma(\pi^*) - \Gamma(\widehat{\pi}) \le{}& |\Gamma(\pi^*) - \widehat{\Gamma}(\pi^*)| \\ &+ |\widehat{\Gamma}(\widehat{\pi}) - \Gamma(\widehat{\pi})|.\end{aligned} \tag{26}$$

For any feasible policy π, Cauchy–Schwarz yields

$$\begin{aligned}|\Gamma(\pi) - \widehat{\Gamma}(\pi)| &\le \frac{V_{\max}}{n}\sum_i \pi_i|\tau_i - \widehat{\tau}_i| \\ &\le V_{\max}\sqrt{\frac{K}{n}}\left\{\frac{1}{n}\sum_i(\tau_i - \widehat{\tau}_i)^2\right\}^{1/2}.\end{aligned} \tag{27}$$

Applying the bound to both terms proves Eq. (18). This calculation assumes that the feasible set, costs, and customer values are the same for the oracle and plug-in policies. It does not supply a confidence interval or resolve causal identification.

B Reproducibility specification

The main analysis uses the exact CSV cited in the data-availability statement. Its SHA-256 digest is

```
16320c9c1ec72448db59aa0a26a0b95401
046bef5d02fd3aeb906448e3055e91
```

The simulation coefficient vectors, ordered according to the eight design columns in Sect. 4, are given in Eq. (28). Line breaks only group vector entries for column-width readability:

$$\begin{aligned}\beta_e^\top &= (-0.40, 0.35, -0.70, 0.50, \\ &\qquad 0.60, 0.90, 0.25, 0.10), \\ \beta_0^\top &= (0.25, -0.35, -0.90, 0.50, \\ &\qquad 0.80, 1.20, -0.20, 0.10), \\ \beta_\Delta^\top &= (0.35, 0, 0.80, -0.50, \\ &\qquad 0, -0.90, 0, 0).\end{aligned} \tag{28}$$

The omitted-covariate models retain only the intercept. Outcome regression estimates the evaluation mean of $\widehat{\mu}_1 - \widehat{\mu}_0$. IPW estimates the mean of $AY/\widehat{e}_\epsilon - (1-A)Y/(1-\widehat{e}_\epsilon)$. The doubly robust estimator averages Eq. (8). Errors are calculated against the exact generating contrast on each evaluation sample, without estimating a survival endpoint.

The executable numerical implementation is embedded in this LaTeX source and supplied separately in Online Resource 1 (`ESM_1.zip`). Compilation extracts `retention_core.py`, `retention_checks.py`, and `retention_validation.py`. Running the validation script executes the core analysis, sensitivity checks, all simulations, and repeated partitions. It requires NumPy and pandas, and downloads the cited CSV if it is absent from the same directory. A digest check verifies the input before analysis. Online Resource 1 additionally contains plotting code, package versions, run instructions, and the complete numerical result tables. No separate bibliography file is required.

Statements and Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Table 8 Fixed implementation and evaluation settings

Setting	Value
Raw sample and analysis cohort	7,043 and 5,517 records
Training / held-out sizes	3,862 / 1,655 records
Joint-label stratification	Technical support × observed retention
Main seed; additional bootstrap seed; simulation seed	20261009; 20261010; 20261011
Cross-fitting	Five training folds, fixed joint-label stratification
Logistic objective	Summed negative log-likelihood + $\frac{1}{2}\sum_{j>0}\beta_j^2$
Logistic numerical stopping	100 iterations or maximum update $< 10^{-7}$
Ridge objective	Summed squared score residuals + $10\sum_{j>0}b_j^2$
Primary propensity clipping	$[0.025, 0.975]$; no clipping active in the main specification
Coverage grid and primary subgroup	51 equally spaced coverage points; top 20%, $n = 331$
Conditional uncertainty	1,000 evaluation-row resamples; fixed fitted models and selected indicators
Interpretation	30 grouped feature permutations; mean absolute prediction change
Average-effect simulation	200 repetitions; 3,000 generated rows, split 2,000 / 1,000
Repeated partitions	30 complete refits; seed 20261012
Decision validation	200 repetitions per mechanism; three mechanisms; seed 20261013
Decision cost ratios	$\{0, 0.05, 0.10, 0.20, 0.30\}$; capacity at most 20%

Competing interests

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Ethics approval and consent to participate

Not applicable. The study analyses a publicly available fictional-company sample and generated data. It involves no recruitment of human participants, direct intervention, or collection of identifiable customer information.

Consent for publication

Not applicable. The article contains no identifiable individual participant information.

Data availability

The 21-column Telco Customer Churn CSV is publicly available from IBM's repository at https://github.com/IBM/telco-customer-churn-on-icp4d/blob/master/data/Telco-Customer-Churn.csv. The exact file is identified by the SHA-256 digest in Appendix B. Generating mechanisms, seeds, and coefficient vectors are specified in Sect. 4 and Appendix B; numerical outputs are supplied in Online Resource 1. No proprietary enterprise records are used.

Code availability

Executable analysis and simulation code is embedded in the submitted LaTeX source and extracted on compilation as `retention_core.py`, `retention_checks.py`, and `retention_validation.py`. The same code is supplied in Online Resource 1 (`ESM_1.zip`). The entry script `retention_validation.py` regenerates customer-level predictions, diagnostic summaries, sensitivity results, repeated partitions, and simulation errors. Online Resource 1 also reproduces the figures. All numerical methods and hyperparameters are specified in the article and source code.

Supplementary information

Online Resource 1: executable analysis and plotting code, software requirements, reproduction instructions, customer-level analysis outputs, and complete sensitivity, partition-stability, and simulation results (`ESM_1.zip`).

Author contributions

Xiaobo Xing: conceptualization, methodology, formal analysis, investigation, data curation, visualization, writing of the original draft, and review

and editing. The use of AI assistance is disclosed in Sect. 4.

Acknowledgements

Not applicable.

References

[1] Barsotti, A., Gianini, G., Mio, C., Lin, J., Babbar, H., Singh, A., Taher, F., Damiani, E.: A decade of churn prediction techniques in the TelCo domain: a survey. SN Computer Science **5**, 404 (2024). https://doi.org/10.1007/s42979-024-02722-7

[2] Ascarza, E.: Retention futility: targeting high-risk customers might be ineffective. Journal of Marketing Research **55**(1), 80–98 (2018). https://doi.org/10.1509/jmr.16.0163

[3] Lemmens, A., Gupta, S.: Managing churn to maximize profits. Marketing Science **39**(5), 956–973 (2020). https://doi.org/10.1287/mksc.2020.1229

[4] Gutierrez, P., Gérardy, J.-Y.: Causal inference and uplift modelling: a review of the literature. Proceedings of Machine Learning Research **67**, 1–13 (2017). https://proceedings.mlr.press/v67/gutierrez17a.html

[5] Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B.: Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences **116**(10), 4156–4165 (2019). https://doi.org/10.1073/pnas.1804597116

[6] Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association **113**(523), 1228–1242 (2018). https://doi.org/10.1080/01621459.2017.1319839

[7] Kennedy, E.H.: Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics **17**(2), 3008–3049 (2023). https://doi.org/10.1214/23-EJS2157

[8] Athey, S., Wager, S.: Policy learning with observational data. Econometrica **89**(1), 133–161 (2021). https://doi.org/10.3982/ECTA15732

[9] Gao, C., Zhang, Z., Yang, S.: Causal customer churn analysis with low-rank tensor block hazard model. Proceedings of Machine Learning Research **235**, 14920–14953 (2024). https://proceedings.mlr.press/v235/gao24q.html

[10] Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019). https://jmlr.org/papers/v20/18-760.html

[11] Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology **66**(5), 688–701 (1974). https://doi.org/10.1037/h0037350

[12] Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983). https://doi.org/10.1093/biomet/70.1.41

[13] Bang, H., Robins, J.M.: Doubly robust estimation in missing data and causal inference models. Biometrics **61**(4), 962–973 (2005). https://doi.org/10.1111/j.1541-0420.2005.00377.x

[14] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J.: Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal **21**(1), C1–C68 (2018). https://doi.org/10.1111/ectj.12097

[15] Kang, J.D.Y., Schafer, J.L.: Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science **22**(4), 523–539 (2007). https://doi.org/10.1214/07-STS227

[16] Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems.

Technometrics **12**(1), 55–67 (1970). https://doi.org/10.1080/00401706.1970.10488634

[17] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)

[18] Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M., Wei, L.J.: Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. Journal of Clinical Oncology **32**(22), 2380–2385 (2014). https://doi.org/10.1200/JCO.2014.55.2208

[19] IBM: Telco Customer Churn, 21-column CSV dataset. IBM public repository. https://github.com/IBM/telco-customer-churn-on-icp4d/blob/master/data/Telco-Customer-Churn.csv. Accessed 9 October 2026

[20] IBM Samples Team: Telco customer churn (11.1.3+). IBM Cognos Analytics Community (2019). https://community.ibm.com/community/user/blogs/steven-macko/2019/07/11/telco-customer-churn-1113. Accessed 9 October 2026

[21] Hunter, J.D.: Matplotlib: a 2D graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55

[22] Austin, P.C.: Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine **28**(25), 3083–3107 (2009). https://doi.org/10.1002/sim.3697

[23] Wilson, E.B.: Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association **22**(158), 209–212 (1927). https://doi.org/10.1080/01621459.1927.10502953