Discover Computing

RESEARCH Open Access

A decision-oriented empirical comparison of predictive and uplift-based scoring under budget constraints

Jianqing Jiang[1], Nor Asilah Wati Abdul Hamid[2], Ng Keng Yap[2] and Choo Wei Chong[3*]

*Correspondence:
Choo Wei Chong
wcchoo@upm.edu.my
[1]Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia (UPM), Serdang, Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM), Serdang, Malaysia
[3]School of Business and Economics (SBE), Universiti Putra Malaysia (UPM), Serdang, Malaysia

Abstract

Decision makers often operate under budget constraints and must allocate limited interventions across large populations. A common approach ranks individuals by predicted response probability and selects the top-k users. However, predictive scoring targets outcome likelihood rather than incremental impact and may allocate resources to individuals who would respond even without intervention. Uplift-based scoring, in contrast, ranks individuals by estimated treatment-induced response and is therefore more directly aligned with intervention allocation. This paper presents a decision-oriented empirical comparison of predictive and uplift-based scoring under a unified budget-constrained top-k targeting framework. Using a large-scale observational digital-lending dataset, we construct a temporally ordered evaluation design in which current-month features and treatment are mapped to next-month drawdown behavior. Predictive models estimate next-month response probability, while uplift-based models estimate incremental treatment effects. All scores are converted into the same top-k policy and evaluated using inverse propensity scoring (IPS), self-normalized weighting (WIS/SNIPS), and doubly robust/augmented inverse probability weighting (DR/AIPW) policy-value estimators. The results show that predictive models provide reasonable response-prediction performance, indicating that they are not weak baselines. Nevertheless, under the unified budget protocol, uplift-based policies achieve higher DR/AIPW policy value than predictive policies across the evaluated budget levels, including when compared against the strongest predictive baseline. IPS estimates are more conservative and sometimes negative, reflecting high weighted control benchmarks among selected users. Supporting analyses using WIS/SNIPS, bootstrap uncertainty, weight trimming, hidden-confounding sensitivity analysis, placebo tests, selected-group decomposition, and response-type diagnostics are consistent with the DR/AIPW-centered findings. These findings provide dataset-specific empirical evidence that uplift-based scoring is better aligned with budget-constrained intervention allocation when the operational objective is incremental impact. The study highlights the importance of evaluating scoring models by the policy value of the decisions they induce, rather than by predictive accuracy alone.

Keywords Budget-constrained targeting, Uplift modeling, Predictive scoring, Policy value, DR/AIPW, Heterogeneous treatment effects, Observational policy evaluation

1 Introduction

Organizations increasingly rely on scoring models to allocate limited interventions across large user populations. In domains such as marketing, recommendation systems, digital platforms, and customer engagement, treatment capacity is often constrained by budget, channel availability, operational cost, or compliance requirements. Under such constraints, the practical question is not whether an intervention is effective on average, but which individuals should be prioritized when only a limited fraction of the population can be targeted.

This decision problem is commonly implemented through a top-k targeting rule. A scoring model ranks individuals, and the decision system selects the highest-ranked users until the budget is exhausted. In this setting, a score does not merely summarize user characteristics; it induces an allocation policy. The quality of the model should therefore be assessed not only by its predictive performance, but also by the value of the decisions generated by the ranking rule [1–4].

A widely used approach is predictive scoring. Predictive models estimate the probability that an individual will produce the target outcome and rank users accordingly. Such models are attractive because they are easy to train, validate, and deploy using standard supervised-learning workflows. Logistic regression, random forests, XGBoost, and LightGBM are commonly used because they are scalable and often provide strong predictive performance. This prediction-oriented logic is also closely related to the broader development of user modeling, recommender systems, customer analytics, and profiling-based personalization [5–10].

However, predictive scoring is not equivalent to intervention allocation. Predictive scores rank users by outcome likelihood, whereas intervention decisions should prioritize users whose outcomes are likely to change because of treatment. From the potential-outcomes perspective, observed responses combine baseline response propensity with treatment-induced change [11–13]. Under budget constraints, this distinction becomes critical. Users with the highest predicted response probabilities may respond even without treatment. Allocating scarce interventions to such users can improve observed response rates while generating limited incremental intervention value.

Uplift modeling and heterogeneous treatment-effect estimation address this issue by focusing on treatment-induced differences rather than outcome levels. Early uplift modeling work introduced the idea of ranking individuals by the incremental effect of treatment rather than by response probability [14–17]. Tree-based uplift models, multi-treatment uplift modeling, and applied uplift studies further extended this logic in marketing, churn prevention, and treatment-allocation settings [18–21]. In parallel, causal machine learning has developed flexible methods for estimating heterogeneous treatment effects, including causal trees, causal forests, generalized random forests, meta-learners, representation-learning approaches, R-learning, and double/debiased machine learning [22–29].

The distinction between predictive and uplift-based scoring is therefore not simply a distinction between two algorithm families. It is a distinction between two decision

signals. Predictive scoring asks who is likely to respond. Uplift-based scoring asks whose response is likely to change because of the intervention. When interventions are costly or capacity-constrained, the second question is more directly aligned with incremental intervention value.

Prior work has developed rich methods for predictive modeling, uplift modeling, causal machine learning, and policy learning. Predictive models are typically evaluated by AUC, calibration, or classification performance. Uplift models are often evaluated using uplift curves, raw AUUC, or Qini. Causal machine-learning methods focus on heterogeneous treatment-effect estimation and related identification or estimation properties. Policy-learning studies evaluate decision rules more directly, but often under settings that differ from common predictive-scoring deployment pipelines [2, 4, 30]. What remains less systematically examined is how predictive and uplift-based scores perform when they are embedded into the same budget-constrained top-k decision protocol and evaluated by the policy value of the decisions they induce.

As shown in Table 1, existing studies primarily focus on predictive accuracy, uplift modeling, treatment-effect estimation, or platform-specific targeting. Fewer studies place predictive and uplift-based scores under the same decision rule and compare the value of the resulting policies. This gap motivates the present study.

Rather than proposing a new causal estimator, this paper provides a decision-oriented empirical comparison of predictive and uplift-based scoring under budget constraints. The central research question is:

Under a unified budget-constrained top-k decision protocol, how do predictive scoring and uplift-based scoring differ in their ability to support effective intervention allocation?

Table 1 Positioning of this study relative to prior scoring and causal-decision approaches

Research stream	Representative studies	Main target and evaluation focus	Remaining gap addressed by this study
Predictive scoring and user modeling	[7–10]	Response probability, user preference, or behavioral propensity; evaluated using AUC, calibration, classification, or recommendation performance	Does not distinguish baseline response propensity from incremental intervention value
Uplift modeling	[14–18, 21]	Incremental response or uplift, $Y(1) - Y(0)$; evaluated using uplift curves, raw AUUC, Qini, or campaign gain	Often not compared with predictive scoring under the same budget-constrained top-*k* decision rule
Causal machine learning and HTE estimation	[22–25, 28, 29]	CATE / ITE; evaluated by treatment-effect estimation, heterogeneity discovery, and nuisance adjustment	Primarily estimation-focused rather than a direct comparison of induced allocation policies
Policy learning and decision evaluation	[1–4, 30]	Policy value or individualized treatment rules; evaluated using offline policy value and treatment-rule performance	Does not directly examine predictive scoring versus uplift-based scoring in the same operational top-*k* targeting protocol
Causal user profiling and response segmentation	[31]	Response segmentation and causal user profiles; evaluated using treatment-response types and uplift-oriented segmentation	Does not provide a unified predictive-versus-uplift policy-value comparison under fixed budget constraints
This study	-	$P(Y_{t+1} = 1 \mid X_t)$ versus $\hat{\tau}(X_t)$; evaluated using DR/AIPW policy value under a unified top-*k* protocol	Provides a decision-oriented empirical comparison of predictive and uplift-based scoring under the same budget-constrained targeting rule

The empirical setting is an observational digital-lending intervention dataset. To strengthen temporal ordering, current-month features and treatment are mapped to next-month drawdown behavior. Predictive models estimate next-month drawdown probability, while uplift-based models estimate the incremental effect of current-month intervention on next-month drawdown. All scores are converted into the same top-k policy and evaluated using a common policy-evaluation framework, with DR/AIPW policy value as the main criterion and IPS, WIS/SNIPS, robustness checks, and response-type diagnostics as supporting evidence.

The paper makes three contributions. First, it reframes the comparison between predictive and uplift-based scoring as a decision problem rather than a pure model-performance comparison. Second, it provides a reproducible empirical protocol that combines temporal outcome construction, leakage prevention, propensity diagnostics, unified top-k policy comparison, DR/AIPW evaluation, robustness checks, and response-type mechanism diagnostics. Third, it provides empirical evidence that predictive models can be reasonable response predictors while still inducing lower policy value than uplift-based policies under budget constraints.

The remainder of the paper is organized as follows. Section 2 defines the budget-constrained targeting problem and the scoring strategies compared in this study. Section 3 describes the data, temporal outcome construction, experimental design, and reproducibility protocol. Section 4 reports the empirical results. Section 5 discusses the implications of the findings, and Sect. 6 concludes.

2 Problem setup and scoring strategies

This section defines the decision problem and the scoring strategies compared in this study. The key distinction is between a score, a policy, and policy value. A score ranks users; a policy converts this ranking into a treatment decision under a budget constraint; policy value evaluates the expected value of the decisions induced by that policy. Throughout the analysis, the decision rule is held fixed and only the scoring function changes. This design allows the comparison to focus on whether response-probability scoring or treatment-effect-oriented scoring is better aligned with budget-constrained intervention allocation.

2.1 Budget-constrained targeting problem

Consider a population of n users indexed by $i = 1, \ldots, n$. At decision period t, the system observes pre-treatment features $X_{i,t}$, a treatment indicator $T_{i,t} \in \{0, 1\}$, and a future binary outcome $Y_{i,t+1}$. In the empirical setting of this paper, $T_{i,t}$ indicates whether the user receives an intervention in the current month, while $Y_{i,t+1}$ indicates whether the same user draws down in the next month.

In each decision round, only a limited fraction of users can be selected for treatment. Let $b \in (0, 1]$ denote the budget ratio and let $k_b = \lfloor bn \rfloor$ denote the corresponding number of selected users. Given a scoring function $s(X_{i,t})$, users are ranked by this score, and the top k_b users are selected. The policy induced by score s at budget b is defined as:

$$\pi_{b,s}(X_{i,t}) = \mathbb{I}\{s(X_{i,t}) \geq c_b\}, \tag{2.1}$$

where c_b is the score threshold corresponding to the top k_b users.

The objective of budget-constrained targeting is not to maximize the observed response rate among selected users. Instead, the objective is to allocate interventions to users for whom treatment is expected to generate additional response. In potential-outcomes terms, this means prioritizing users with larger values of $Y_{i,t+1}(1) - Y_{i,t+1}(0)$, although both potential outcomes cannot be observed for the same user. This unobservability motivates the use of estimated treatment-effect scores.

2.2 Predictive scoring

Predictive scoring ranks users by their estimated probability of the target outcome. In this study, predictive models estimate next-month drawdown probability:

$$s_{\text{pred}}(X_{i,t}) = \widehat{P}(Y_{i,t+1} = 1 \mid X_{i,t}). \tag{2.2}$$

This score answers the question: which users are most likely to draw down next month?

Predictive scoring is useful for forecasting and monitoring because it directly estimates expected response probability. It is also compatible with standard supervised-learning evaluation, including AUC and calibration. However, predictive scoring does not distinguish baseline response propensity from treatment-induced change. A user with a high predicted probability of next-month drawdown may have a high probability of drawing down regardless of whether treatment is delivered. Under a limited intervention budget, ranking such users highly may allocate treatment to users with limited incremental intervention value.

2.3 Uplift-based scoring

Uplift-based scoring ranks users by estimated incremental response. In this study, the uplift score is defined as:

$$s_{\text{uplift}}(X_{i,t}) = \widehat{\tau}(X_{i,t}). \tag{2.3}$$

The score estimates the conditional treatment effect, $E[Y_{i,t+1}(1) - Y_{i,t+1}(0) \mid X_{i,t}]$. When implemented through potential-outcome probability models, it can be represented as the difference between estimated treated and control response probabilities.

This score answers a different question: which users are most likely to change their next-month behavior because of the current-month intervention?

The distinction between predictive and uplift-based scoring is therefore not only algorithmic. It is a distinction between two decision signals. Predictive scoring ranks users by expected outcome level, whereas uplift-based scoring ranks users by estimated outcome change. When interventions are costly or capacity-constrained, the latter is more directly aligned with the objective of maximizing incremental intervention value.

2.4 Unified policy comparison

A fair comparison requires predictive and uplift-based scores to be evaluated under the same deployment rule. This study therefore converts each score into the same top-k policy defined in Eq. (2.1). The data structure, feature space, temporal split, budget levels, and decision rule are held constant; only the scoring function used to rank users changes.

For each score s and budget b, the induced policy is $\pi_{b,s}$, and the corresponding policy value is denoted by $V(\pi_{b,s})$. The main comparison is:

$$\Delta_b = \widehat{V}(\pi_{b,\text{uplift}}) - \widehat{V}(\pi_{b,\text{pred}}), \tag{2.4}$$

where $\widehat{V}(\cdot)$ denotes an offline policy-value estimator. A positive Δ_b indicates that the uplift-based policy achieves higher estimated policy value than the predictive policy under the same budget constraint.

This formulation separates model scoring from policy evaluation. Predictive AUC and calibration are reported to confirm that predictive baselines are not weak. Raw AUUC and Qini are used only as auxiliary uplift-ranking diagnostics. The main empirical comparison is conducted at the policy level using the common evaluation framework described in Sect. 3.

Because the study uses observational data, policy-value interpretation depends on standard causal assumptions, including consistency, conditional exchangeability, and positivity. These assumptions are addressed empirically through temporal ordering, pre-treatment feature construction, propensity diagnostics, covariate balance checks, DR/AIPW estimation, and sensitivity analyses in the following sections.

3 Data, experimental design, and reproducibility

This section describes the empirical design used to compare predictive and uplift-based scoring under a unified budget-constrained targeting protocol. The purpose is to make the comparison reproducible and decision-oriented. The design follows four principles. First, treatment, covariates, and outcomes are temporally ordered to avoid same-period leakage. Second, all scoring strategies use the same pre-treatment feature space and the same train-validation-evaluation split. Third, all scores are converted into the same top-k decision rule. Fourth, policy value is evaluated using inverse-propensity, self-normalized, and doubly robust estimators.

The empirical structure is:

$$X_{i,t}, T_{i,t} \rightarrow Y_{i,t+1}, \tag{3.1}$$

where $X_{i,t}$ denotes user characteristics and historical behaviors available at the current decision month, $T_{i,t}$ denotes current-month treatment assignment, and $Y_{i,t+1}$ denotes next-month drawdown.

3.1 Data source and temporal outcome construction

The empirical analysis uses a large-scale monthly user-level observational dataset from a digital lending and user engagement context. Each observation corresponds to a user-month decision instance and contains user characteristics, historical behavioral variables, treatment assignment, and drawdown behavior.

The raw panel contains 907,893 user-month observations from 184,668 users over six monthly periods. The original outcome variable records whether a user made a drawdown in the current month. However, because both treatment and drawdown are observed at monthly frequency, using current-month drawdown as the outcome for current-month treatment would create ambiguity in temporal ordering. Without exact

intra-month timestamps, it is not possible to verify whether treatment always occurred before drawdown within the same month.

To avoid same-period leakage and reverse-causality concerns, this study defines the outcome as next-month drawdown. For each user-month observation in period t, the same user is linked to period $t+1$, and the next-month drawdown indicator is used as the outcome:

$$Y_{i,t+1} = \text{Drawdown}_{i,t+1}. \tag{3.2}$$

The treatment remains the current-month intervention:

$$T_{i,t} = \text{Treatment}_{i,t}. \tag{3.3}$$

Thus, the empirical estimand is the effect of current-month intervention on next-month drawdown behavior.

3.2 Ethics and data governance

This study was based on retrospective, de-identified secondary operational data provided by Haier Consumer Finance Co., Ltd. for academic research purposes. The dataset was generated from routine platform operations before the research analysis was conducted. The authors did not recruit participants, collect biological materials or human tissue samples, conduct prospective human-subject experiments, directly administer interventions, or collect new personal data for this study.

Before being provided to the authors, the dataset was de-identified by the data provider. The authors did not access names, identity numbers, telephone numbers, addresses, bank account information, or other directly identifiable personal information. All analyses were conducted on de-identified user-month records, and all reported results are presented only in aggregated or derived analytical form.

The study was conducted in accordance with applicable institutional requirements, data-provider restrictions, confidentiality obligations, and relevant regulations governing the use of de-identified secondary operational data.

3.3 Analytical sample and temporal split

After mapping next-month outcomes, the final analytical sample contains 721,861 user-month observations. The last observed month is excluded because its next-month outcome is unavailable.

The sample is split using an out-of-time design rather than random sampling. Earlier months are used for training, the following month is used for validation, and the most recent month with observable next-month outcomes is used for final evaluation (Table 2).

Table 2 Analytical sample construction and temporal split

Split	Periods	Observations	Purpose
Training set	M1-M3	417,063	Model fitting
Validation set	M4	150,347	Model and policy selection
Evaluation set	M5	154,451	Final policy evaluation
Excluded period	M6	–	Excluded because next-month outcome is unavailable

The evaluation set is not used for feature selection, model fitting, hyperparameter tuning, propensity-model selection, or policy selection. This temporal design approximates prospective deployment and reduces the risk of information leakage.

3.4 Feature construction and leakage prevention

All models use the same pre-treatment feature space. Features are constructed from user attributes, credit status, risk history, borrowing history, utilization behavior, and historical intervention exposure. Current-month outcomes, future variables, pre-computed treatment-effect scores, clustering labels, and response-type labels are excluded from model inputs.

The feature groups are summarized in Table 3.

The same feature matrix is used for predictive, uplift, propensity, and outcome-regression models. Therefore, differences in policy value can be attributed to the scoring objective rather than differences in feature availability. A full variable dictionary is included in the reproducibility materials where data-sharing approval permits.

3.5 Scoring models and model selection

This study compares predictive-response scoring and uplift-based treatment-effect scoring. The two scoring families are trained under the same temporal split and feature space, but they estimate different target quantities.

3.5.1 Predictive scoring

Predictive models estimate the next-month drawdown score defined in Eq. (2.2). These models answer the question of who is most likely to draw down next month.

The predictive model set includes logistic regression, random forest, XGBoost, and LightGBM. Predictive model quality is reported using conventional response-prediction diagnostics, including AUC and calibration. These diagnostics are used to show that predictive baselines are not weak. They are not used as the main decision metric.

3.5.2 Uplift-based scoring

Uplift-based models estimate the treatment-effect-oriented score defined in Eq. (2.3). These models answer the question of whose next-month drawdown behavior is most likely to change because of current-month treatment.

Table 3 Feature groups used for predictive and uplift-based scoring

Feature group	Examples	Interpretation	Timing
Demographic attributes	Age, gender, education, marital status	Basic user characteristics	Pre-treatment
Credit status	Credit limit, account age	Credit capacity and account relationship	Pre-treatment
Risk history	Historical overdue indicators and overdue counts	Prior risk behavior	Pre-treatment
Borrowing history	Prior drawdown, loan, and rejection counts	Historical borrowing behavior	Pre-treatment
Utilization intensity	Credit-normalized historical principal ratios	Relative credit utilization	Pre-treatment
Historical intervention exposure	Prior channel-specific exposure indicators	Previous contact or intervention history	Pre-treatment

The uplift model set includes S-, T-, X-, R-, and DR-learner specifications. Raw AUUC and Qini are reported as auxiliary uplift-ranking diagnostics. They are not treated as the main causal policy-value criteria. The main policy comparison is conducted using DR/AIPW policy value after each score is converted into the same top-k decision rule.

3.5.3 Model roles in the empirical design

Table 4 summarizes the role of each model family in the empirical design.

All models use fixed random seeds, identical training and validation splits, and the same pre-treatment feature matrix. Candidate scores are selected before final evaluation. The M5 evaluation set is used only for reporting final policy-value estimates.

3.6 Unified Top-k budget protocol

All scores are converted into the top-k policy defined in Eq. (2.1). For each evaluated budget ratio, users in the evaluation set are ranked by the corresponding score, and the highest-ranked users are selected until the budget is exhausted.

The evaluated budget ratios are 0.5%, 1%, 2%, 3%, 5%, 10%, 20%, and 30%. The decision rule is identical across all models; only the scoring function differs. This ensures that performance differences reflect differences in scoring objectives rather than differences in deployment rules.

3.7 Propensity score estimation, causal assumptions, and diagnostics

Because treatment assignment is observational, policy evaluation requires adjustment for treatment-selection bias. Let

$$e(X_{i,t}) = P(T_{i,t} = 1 \mid X_{i,t}) \tag{3.4}$$

denote the propensity score.

Random forest is used as the main propensity model because it provides stronger empirical overlap and more stable inverse-propensity weights in this dataset. Logistic regression is used as a robustness specification.

The policy-value interpretation relies on standard causal assumptions. Consistency requires that the observed outcome equals the potential outcome under the observed treatment. Conditional exchangeability requires that, after conditioning on observed pre-treatment covariates, treatment assignment is independent of potential outcomes.

Table 4 Model families and empirical roles

Predictive scoring	LR/RF/XGBoost/LightGBM	Estimate next-month response probability	Reference predictive baselines
Main predictive baseline	LightGBM	Strong predictive score for policy comparison	Compared with uplift policy under top-k
Uplift scoring	S-, T-, X-, R-, DR-learners	Estimate treatment-induced response differences	Candidate uplift scores
Main uplift policy	R-learner	Main uplift score used for policy comparison	Compared with predictive policy under top-k
Propensity model	Random forest	Estimate treatment assignment probability	Main adjustment model
Propensity robustness	Logistic regression	Alternative propensity specification	Robustness check
Outcome nuisance models	Treated and control outcome regressions	Estimate $\hat{\mu}_1(X)$ and $\hat{\mu}_0(X)$	DR/AIPW policy-value estimation

Positivity requires that each user has a nonzero probability of receiving either treatment condition:

$$0 < e(X_{i,t}) < 1. \tag{3.5}$$

These assumptions cannot be proven from observational data. The study therefore reports diagnostics that assess their empirical plausibility. Propensity diagnostics include overlap visualization, tail behavior, inverse-propensity weight diagnostics, and effective sample size for treated and control users. Effective sample size is computed as:

$$ESS = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}. \tag{3.6}$$

These diagnostics assess common support and the stability of inverse-propensity weighting. They support, but do not prove, the conditional exchangeability and positivity assumptions.

3.8 Offline policy evaluation

The main empirical objective is to compare the policy value induced by predictive and uplift-based scoring under the same budget constraint. The top-k policies are evaluated using inverse-propensity, self-normalized, and doubly robust estimators. For notational simplicity, this subsection writes $X_i = X_{i,t}$, $T_i = T_{i,t}$, and $Y_i = Y_{i,t+1}$.

3.8.1 IPS estimator

For a policy $\pi(X_i)$, the inverse propensity score estimator is:

$$\widehat{V}_{IPS}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\left[\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}\right]. \tag{3.7}$$

This estimator compares the weighted treated outcome with the weighted control benchmark among selected users. Since this quantity is an incremental policy-value estimate rather than a raw response rate, negative values may occur. IPS is therefore interpreted as a conservative auxiliary estimator rather than the sole basis for inference.

3.8.2 WIS/SNIPS estimator

To reduce sensitivity to extreme weights, a self-normalized estimator is also reported:

$$\widehat{V}_{SNIPS}(\pi) = \frac{\sum_i \pi(X_i)\frac{T_i Y_i}{\hat{e}(X_i)}}{\sum_i \pi(X_i)\frac{T_i}{\hat{e}(X_i)}} - \frac{\sum_i \pi(X_i)\frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}}{\sum_i \pi(X_i)\frac{(1-T_i)}{1-\hat{e}(X_i)}}. \tag{3.8}$$

This estimator is used as a robustness check for IPS and is reported together with the main policy-value results.

3.8.3 DR/AIPW estimator

Let $\hat{\mu}_1(X_i)$ and $\hat{\mu}_0(X_i)$ denote estimated treated and control outcome regressions. The doubly robust policy-value estimator is:

$$\widehat{V}_{DR}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)} - \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)}\right]. \tag{3.9}$$

DR/AIPW is emphasized as the main policy-value estimator because it combines propensity weighting with outcome regression. This does not eliminate all causal identification concerns, but it provides a stronger basis for policy-value comparison than IPS alone.

For each budget b, the policy-value difference follows Eq. (2.4). A positive Δ_b indicates that the uplift-based policy achieves higher estimated incremental policy value under the same budget.

3.9 Robustness and sensitivity analyses

Several additional analyses are conducted to assess the robustness of the policy-value comparison.

First, bootstrap confidence intervals are computed by resampling the evaluation set while holding the fitted models fixed. This estimates uncertainty in the evaluation sample rather than full model re-estimation variability.

Second, weight-trimming and propensity-clipping analyses are conducted to test whether policy-value differences are driven by extreme inverse-propensity weights.

Third, propensity-model robustness is assessed by comparing the main random-forest propensity model with a logistic-regression propensity specification.

Fourth, hidden-confounding sensitivity is evaluated by perturbing the treatment-assignment odds using a sensitivity factor Γ:

$$\mathrm{odds}_{\Gamma}(X_i) = \Gamma \cdot \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}. \tag{3.10}$$

The corresponding perturbed propensity score is:

$$\hat{e}_{\Gamma}(X_i) = \frac{\mathrm{odds}_{\Gamma}(X_i)}{1+\mathrm{odds}_{\Gamma}(X_i)}. \tag{3.11}$$

The evaluated sensitivity levels are $\Gamma = 1.00, 1.10, 1.25, 1.50$, and 2.00. This analysis does not eliminate hidden-confounding concerns, but it evaluates whether the main policy-ranking conclusion remains stable under deviations from the estimated treatment-assignment mechanism.

Finally, placebo tests are conducted by randomly permuting the uplift score and repeating the top-k evaluation. The purpose is to verify that the observed uplift advantage is not mechanically produced by the evaluation procedure.

3.10 Response-type mechanism diagnostics

To interpret why predictive and uplift-based policies select different users, response-type mechanism diagnostics are conducted using estimated potential-outcome probabilities. Let

$$\hat{\mu}_1(X_{i,t}) = \widehat{P}(Y_{i,t+1} = 1 \mid X_{i,t}, T_{i,t} = 1) \tag{3.12}$$

and

$$\hat{\mu}_0(X_{i,t}) = \widehat{P}(Y_{i,t+1} = 1 \mid X_{i,t}, T_{i,t} = 0) \tag{3.13}$$

denote the estimated treated and control outcome probabilities for next-month drawdown.

Because true potential outcomes $Y_i(1)$ and $Y_i(0)$ are never jointly observed, response types cannot be directly observed. Proxy response types are therefore constructed by thresholding the estimated potential-outcome probabilities:

$$\widehat{Y}_i(1) = \mathbb{I}\{\hat{\mu}_1(X_{i,t}) \geq \theta\}, \tag{3.14}$$

$$\widehat{Y}_i(0) = \mathbb{I}\{\hat{\mu}_0(X_{i,t}) \geq \theta\}, \tag{3.15}$$

where θ is fixed before final evaluation and held constant across policies.

The four response-type proxies are defined as follows (Table 5).

These response-type labels are used only for post-hoc mechanism interpretation. They are not used to train predictive models, estimate uplift scores, fit propensity models, or construct top-k policies. Therefore, they do not affect the policy-value estimates.

Because the labels are inferred from estimated potential-outcome probabilities, they should not be interpreted as externally observed causal ground truth. Instead, they provide a mechanism-oriented diagnostic of how predictive and uplift-based policies allocate treatment budgets across users with different estimated response patterns.

3.11 Reproducibility materials

To support reproducibility, this study reports the experimental protocol, model settings, random seeds, evaluation formulas, and generated result tables in the manuscript and supplementary materials where applicable. Because the original operational data are subject to confidentiality, commercial restrictions, and data-provider requirements, raw user-level data are not publicly available.

De-identified analytical materials, non-sensitive variable descriptions, or code excerpts may be made available from the corresponding author upon reasonable request and subject to approval by the data provider and relevant institutional requirements. These materials are intended to support protocol-level and code-level verification where permitted, but they do not include raw operational data or personally identifiable information.

4 Empirical results

This section evaluates whether predictive scoring and uplift-based scoring lead to different intervention-allocation decisions under the same budget-constrained top-k protocol. The comparison is conducted on the out-of-time evaluation set, using current-month covariates and treatment exposure to evaluate next-month drawdown. The empirical focus is therefore not only whether a model predicts response accurately, but whether

Table 5 Inferred response-type definitions

Response type	Proxy definition	Interpretation
Persuadable	$\widehat{Y}(1) = 1, \widehat{Y}(0) = 0$	Likely to respond because of treatment
Sure thing	$\widehat{Y}(1) = 1, \widehat{Y}(0) = 1$	Likely to respond even without treatment
Lost cause	$\widehat{Y}(1) = 0, \widehat{Y}(0) = 0$	Unlikely to respond under either condition
Do-not-disturb	$\widehat{Y}(1) = 0, \widehat{Y}(0) = 1$	Treatment may reduce response probability

the score it produces induces a targeting policy with higher incremental intervention value.

Figure 1 summarizes the empirical workflow. The analysis begins with the temporal construction of treatment and outcome variables, followed by propensity and covariate-balance diagnostics. Predictive performance and uplift-ranking diagnostics are then examined as reference evidence before the induced top-k policies are compared using IPS, WIS/SNIPS, and DR/AIPW policy value. Because treatment assignment is observational, DR/AIPW is treated as the main policy-value estimator, while IPS and raw AUUC/Qini are interpreted as auxiliary diagnostics.

4.1 Data construction and treatment-outcome distribution

The final analytical sample is constructed by linking each user-month observation at month t to the same user's drawdown status in month $t+1$. The last observed month is excluded because the corresponding next-month outcome is unavailable. This design aligns current-month covariates and treatment exposure with next-month drawdown, thereby reducing the risk of same-period leakage.

The temporal split is summarized in Table 6. M1-M3 are used for training, M4 for validation, and M5 for out-of-time evaluation, while M6 is excluded from the outcome construction. Table 7 shows the monthly treatment and outcome distributions. The variation in treatment rates and next-month drawdown rates across months supports the use of a temporal evaluation protocol rather than a random split. Figure 2 illustrates the current-month intervention to next-month outcome design.

The resulting sample implements the intended treatment-outcome structure: current-month features and intervention exposure, (X_t, T_t), are used to evaluate next-month drawdown, Y_{t+1}. This construction ensures that the evaluation target is temporally aligned with the intervention decision.

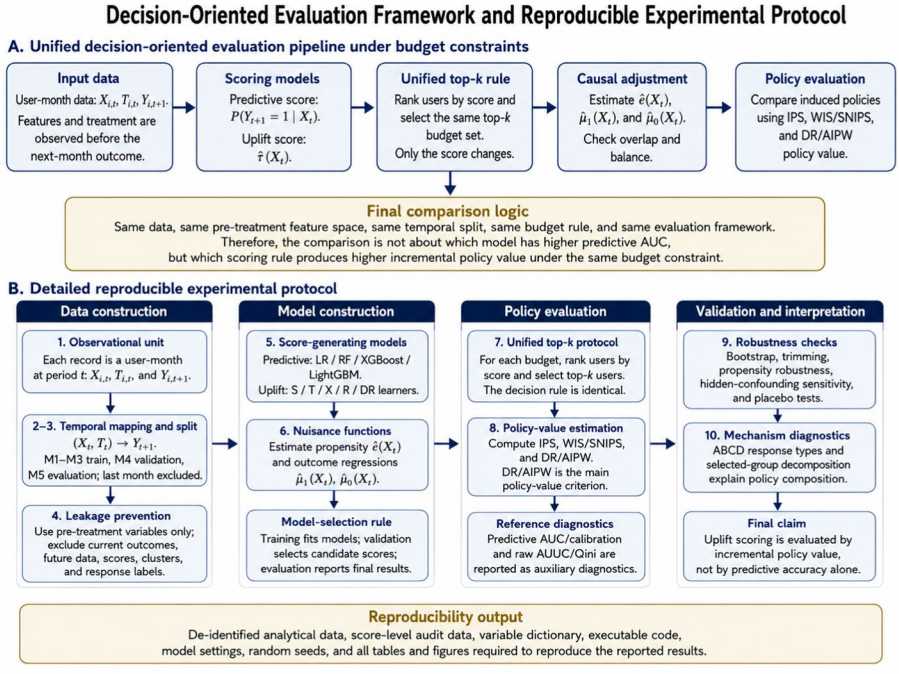

Fig. 1 Experimental roadmap and evidence chain for Sect. 4

Table 6 Analytical sample construction and temporal split

Sample	Obs.	Users	T rate	Y rate	Notes
Final analytical sample	721,861	169,793	36.7%	52.1%	M1-M5 after next-month outcome mapping; M6 excluded
Training set	417,063	150,223	38.3%	51.7%	M1-M3
Validation set	150,347	150,347	29.3%	53.5%	M4
Evaluation set	154,451	154,451	39.9%	51.8%	M5

Table 7 Monthly treatment and next-month outcome distribution

Period	Split	Observations	Treatment rate	Next-month outcome rate	Current-month drawdown rate
M1	Train	135,002	40.6%	50.2%	51.6%
M2	Train	137,635	35.9%	52.3%	51.2%
M3	Train	144,426	38.4%	52.4%	53.8%
M4	Validation	150,347	29.3%	53.5%	53.4%
M5	Evaluation	154,451	39.9%	51.8%	53.8%

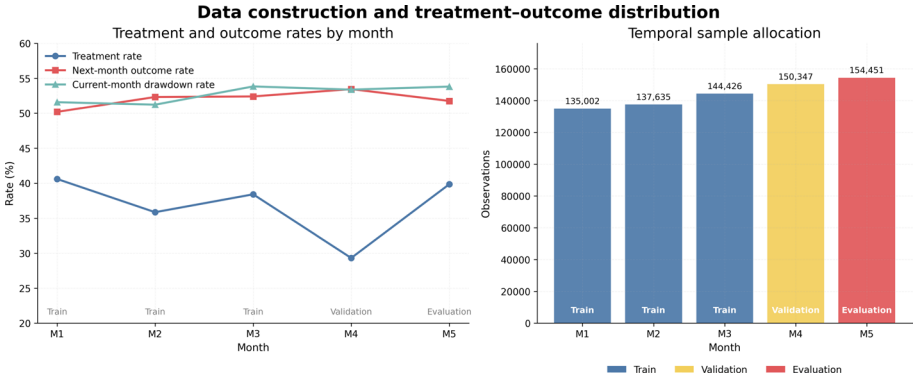

Fig. 2 Monthly treatment rate and next-month outcome rate

Table 8 Propensity score and inverse-weight diagnostics

Propensity model	Treatment AUC	p01 e(X)	Median e(X)	p99 e(X)	p99 weight	Max weight	ESS treated	ESS control
RF main	0.889	0.049	0.25	0.932	7.055	30.093	38,550	62790.5
Logistic robustness	0.909	0.01	0.304	0.99	24.695	100	16016.1	6426.24

4.2 Propensity diagnostics and covariate balance

Since treatment assignment is observational, propensity diagnostics and covariate-balance checks are examined before comparing policy value. The random-forest propensity model is used as the main treatment-assignment model, with logistic regression retained as a robustness specification. This distinction matters because the logistic model produces more extreme inverse-propensity weights, whereas the random-forest model provides more stable common-support behavior.

Propensity-score and inverse-weight diagnostics are reported in Table 8, and covariate balance before and after IPW adjustment is summarized in Table 9. Figure 3 shows the

Table 9 Covariate balance summary before and after IPW

Diagnostic	Before IPW	After IPW
No. of features	32	32
Mean absolute SMD	0.268	0.133
Maximum absolute SMD	1.266	0.770
Share of features with absolute SMD < 0.10	37.5%	59.4%

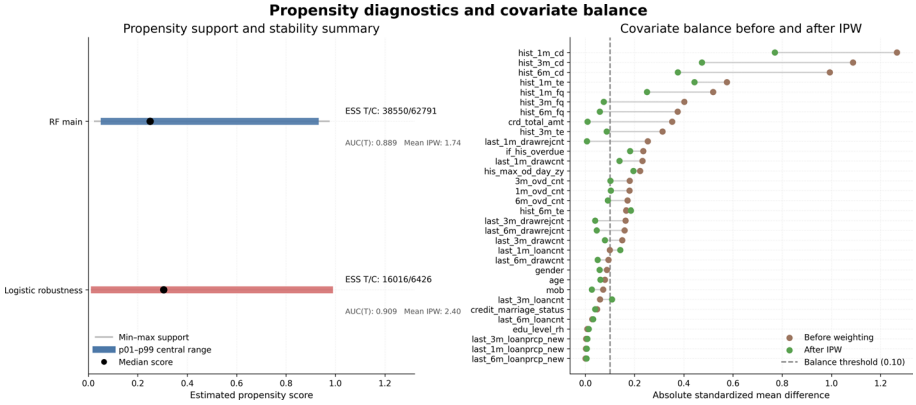

Fig. 3 Propensity overlap and covariate balance diagnostics

Table 10 Predictive response-model reference metrics

Predictive model	AUC	PR-AUC	Brier score	Mean score	Observed outcome rate
predictive_lgbm	0.6705	0.6586	0.2353	0.5198	0.5177
predictive_rf	0.6556	0.6474	0.2323	0.5341	0.5177
predictive_xgboost	0.6523	0.639	0.2396	0.5173	0.5177
predictive_lr	0.6508	0.6547	0.2339	0.4888	0.5177

propensity-score distribution and common-support pattern under the main propensity specification.

The diagnostics indicate that the random-forest propensity model provides usable overlap and substantially more stable weights than the logistic specification. Covariate balance improves after IPW adjustment, although some residual imbalance remains. For this reason, the main policy comparison is not based on IPS alone. DR/AIPW is used as the primary policy-value estimator, and hidden-confounding sensitivity analysis is reported later to assess the robustness of the decision-oriented conclusion.

4.3 Predictive performance reference: AUC and calibration

Predictive models are evaluated first as next-month response models. These results are not used as the primary decision criterion, but they establish whether the predictive baselines provide meaningful response-probability scores. This matters because the central comparison is not between a strong uplift model and a weak predictive model, but between two different scoring objectives under the same budget-constrained targeting rule.

Table 10 reports the predictive performance of LR, RF, XGBoost, and LightGBM, while Figure 4 presents the corresponding performance and calibration-related diagnostics.

The predictive baselines achieve reasonable discrimination for next-month response prediction, with AUC values around 0.65–0.67. LightGBM performs best among the

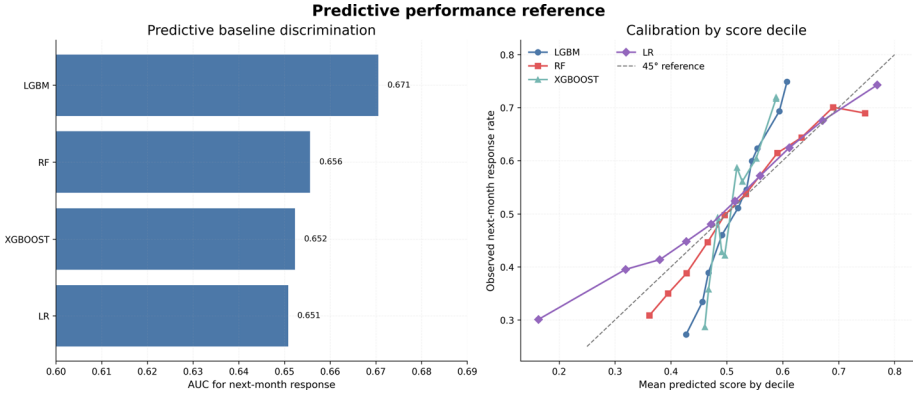

Fig. 4 Predictive AUC reference and score distributions

predictive models in this reference comparison, while RF, XGBoost, and LR also provide non-trivial predictive performance. These results show that the subsequent policy-value comparison is not driven by artificially weak predictive baselines.

4.4 Uplift ranking diagnostics: AUUC, Qini, and decision-value alignment

Uplift scores serve a different purpose from predictive response scores. They are intended to rank users by estimated incremental response under treatment relative to no treatment, rather than by the probability of observing a response. For this reason, uplift-ranking diagnostics are reported separately from predictive AUC.

Before interpreting these diagnostics, the ranking direction is checked by comparing the original descending-score order with its reverse ordering. This check helps ensure that the main positive-AUUC candidates are not artifacts of an inverted sorting convention. Because treatment assignment is observational rather than randomized, however, raw AUUC and Qini curves are used only as auxiliary uplift-ranking diagnostics. They are not treated as the main basis for causal policy comparison.

The raw AUUC is computed as the trapezoidal area under the cumulative uplift curve:

$$U(p) = \bar{Y}_1^{(p)} - \bar{Y}_0^{(p)}. \tag{4.1}$$

Here, $\bar{Y}_1^{(p)}$ and $\bar{Y}_0^{(p)}$ denote cumulative treated and control response rates within the top-p fraction ranked by a given score. The corresponding Qini curve is computed as:

$$Q(p) = Y_1^{(p)} - Y_0^{(p)} \frac{N_1^{(p)}}{N_0^{(p)}}. \tag{4.2}$$

Because these curves are unadjusted, they are interpreted as ranking diagnostics rather than definitive causal evidence.

Table 11 reports the uplift-ranking diagnostics across candidate uplift scores, and Figure 5 shows the corresponding AUUC/Qini behavior.

Among the original candidate uplift scores, the T-Learner with LightGBM shows the strongest auxiliary raw-AUUC signal, while R- and DR-learner variants provide stable positive DR/AIPW policy value. The nonlinear R-Learner capacity check suggests that a LightGBM treatment-effect stage may further improve both ranking and policy-value performance. Since this extension is exploratory rather than part of the pre-specified

Table 11 Uplift-ranking and decision-value diagnostics

Uplift score	Qini area	Raw AUUC area	Direction check	DR@10%
t_lgbm	−0.0033	0.0563	Normal	0.0118
dr_linear	−0.0045	0.0262	Normal	0.0129
r_linear	−0.0043	0.0208	Normal	0.0123
r_lgbm (LGBM capacity check)	0.0015	0.0564	Normal	0.0166
dr_lgbm (LGBM capacity check)	−0.0003	0.0394	Normal	0.0093
t_rf	−0.0095	−0.0257	Normal	0.0010
x_linear	−0.0206	−0.0592	Reverse	0.0095
s_lgbm	−0.0135	−0.0673	Normal	−0.0013

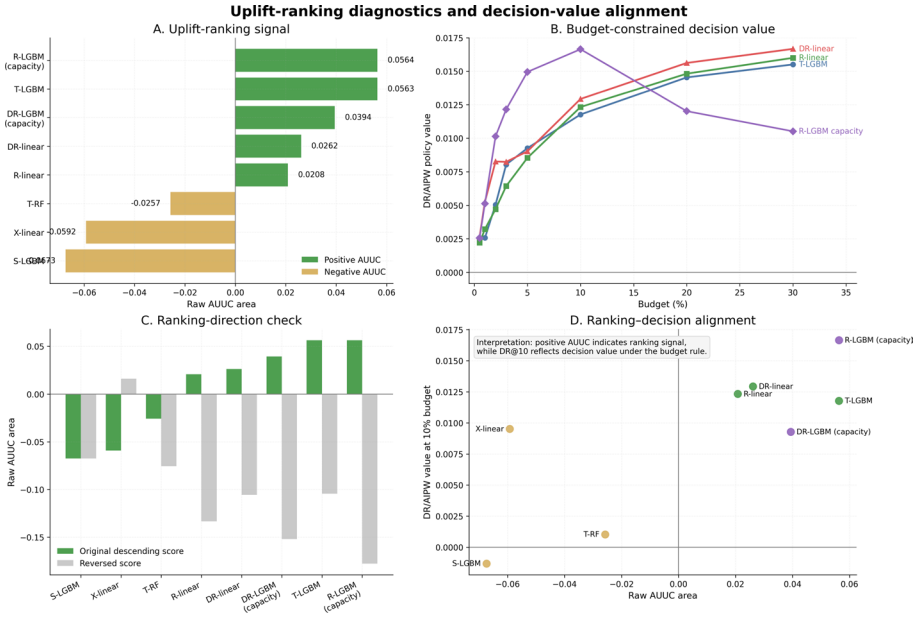

Fig. 5 Uplift-ranking diagnostics and decision-value alignment

main comparison, the main policy analysis remains anchored on the validation-selected R-learner uplift score.

4.5 Main policy value comparison: standard predictive baseline

The first policy-value comparison uses LightGBM as the standard predictive baseline and the validation-selected R-learner score as the main uplift policy. Both scores are translated into the same top-k decision rule: for each budget level, users are ranked by the corresponding score and the top-k users are selected for intervention.

Table 12 reports IPS, WIS/SNIPS, and DR/AIPW policy-value comparisons between the LightGBM predictive policy and the R-learner uplift policy. The differences across evaluated budget levels are shown in Fig. 6.

Under the unified top-k protocol, the uplift-based policy achieves higher DR/AIPW policy value than the LightGBM predictive policy across all evaluated budget levels. IPS differences are mostly positive, although they become small or mixed at several cutoffs. This pattern is consistent with the construction of IPS, which compares selected treated outcomes against a propensity-weighted control benchmark rather than against raw response rates. IPS is therefore interpreted as a conservative auxiliary estimator, while DR/AIPW provides the main policy-value evidence.

Table 12 Main comparison against the standard predictive baseline

Budget	Selected N	Predictive IPS	Uplift IPS	Diff. IPS	Predictive DR	Uplift DR	Diff. DR
0.5%	772	−0.0025	−0.0004	0.0021	−0.0006	0.0022	0.0029
1.0%	1544	−0.004	−0.0031	0.001	−0.001	0.0032	0.0042
2.0%	3089	−0.0066	−0.0075	−0.001	−0.0014	0.0047	0.0061
3.0%	4633	−0.0106	−0.0106	−0.0001	−0.0026	0.0064	0.009
5.0%	7722	−0.0198	−0.0134	0.0064	−0.0047	0.0085	0.0133
10.0%	15,445	−0.0369	−0.0192	0.0178	−0.0095	0.0123	0.0218
20.0%	30,890	−0.0413	−0.0253	0.0161	−0.0108	0.0148	0.0256
30.0%	46,335	−0.0505	−0.051	−0.0004	−0.0125	0.016	0.0285

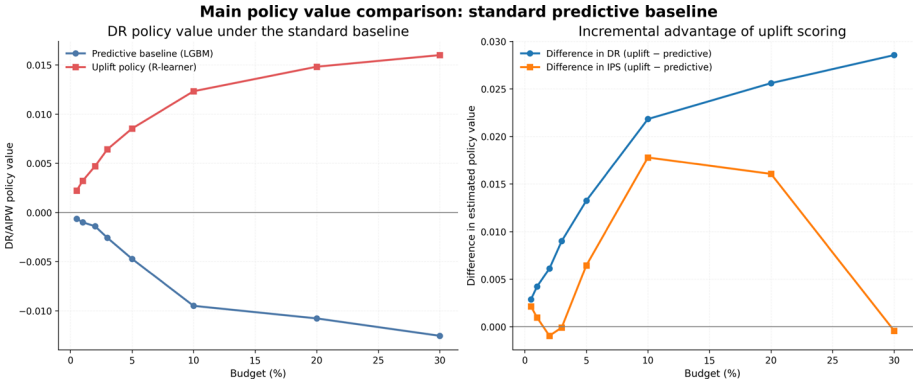

Fig. 6 Standard predictive baseline comparison

Table 13 Robustness comparison against the strongest predictive baseline

Budget	Selected N	Predictive IPS	Uplift IPS	Diff. IPS	Predictive DR	Uplift DR	Diff. DR
0.5%	772	−0.0015	−0.0004	0.0012	−0.0005	0.0022	0.0027
1.0%	1544	−0.003	−0.0031	−0.0001	−0.0011	0.0032	0.0043
2.0%	3089	−0.0057	−0.0075	−0.0018	−0.0019	0.0047	0.0066
3.0%	4633	−0.0068	−0.0106	−0.0038	−0.0024	0.0064	0.0089
5.0%	7722	−0.0098	−0.0134	−0.0036	−0.0031	0.0085	0.0116
10.0%	15,445	−0.011	−0.0192	−0.0082	−0.0033	0.0123	0.0156
20.0%	30,890	−0.0341	−0.0253	0.0088	−0.0078	0.0148	0.0227
30.0%	46,335	−0.0463	−0.051	−0.0046	−0.0076	0.016	0.0236

4.6 Robustness against the strongest predictive baseline

The analysis also compares the uplift policy with the strongest predictive baseline selected from validation performance. This additional comparison is useful because predictive models are expected to perform well when the objective is to rank users by response probability.

The robustness comparison is reported in Table 13, and the corresponding policy-value differences across budget levels are shown in Fig. 7.

The uplift policy continues to achieve positive DR/AIPW differences across all evaluated budget levels when compared with the stronger predictive baseline. IPS differences remain mixed, again suggesting that IPS should not be interpreted in isolation. The comparison supports the decision-oriented interpretation of the results: even when predictive baselines are competitive for next-month response prediction, their ranking objective is not necessarily aligned with incremental intervention value.

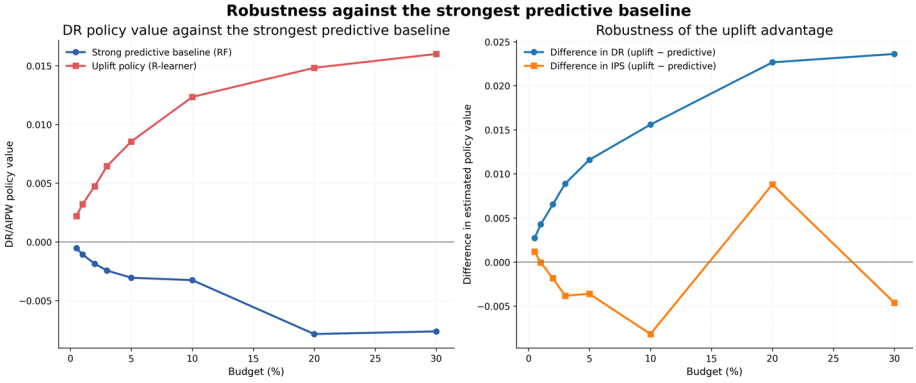

Fig. 7 Strong predictive baseline comparison

Table 14 All-model policy-value comparison at 10% budget

Score policy	IPS value	DR value
uplift_dr_linear	−0.0343	0.0129
uplift_r_linear	−0.0192	0.0123
uplift_t_lgbm	−0.0135	0.0118
uplift_x_linear	−0.005	0.0095
uplift_t_rf	−0.0209	0.001
uplift_s_lgbm	−0.0126	−0.0013
predictive_rf	−0.011	−0.0033
predictive_xgboost	−0.02	−0.0058
predictive_lr	−0.0214	−0.0073
predictive_lgbm	−0.0369	−0.0095

4.7 All-model comparison at 10% budget

Table 14 compares all predictive and uplift policies at the representative 10% budget level under the same evaluation rule.

At the 10% budget level, the highest DR/AIPW policy values are achieved by uplift-family models, especially DR-, R-, T-, and X-learner variants. Predictive models show lower DR/AIPW policy values despite their reasonable predictive AUC in Sect. 4.3. This cross-model comparison reinforces the distinction between predicting who is likely to respond and identifying whose response is most likely to be changed by intervention.

4.8 Bootstrap, WIS/SNIPS, and weight-trimming robustness

Bootstrap uncertainty, self-normalized weighting, and weight-trimming analyses are used to examine whether the main policy-value pattern is driven by sampling variability or extreme inverse-propensity weights.

Bootstrap uncertainty for the main policy-value differences is reported in Table 15. WIS/SNIPS robustness checks are shown in Table 16, and the weight-trimming analysis is reported in Table 17. Figure 8 summarizes the robustness behavior of the main DR/AIPW policy-value comparison.

The bootstrap results show that DR/AIPW differences remain positive at the evaluated budget levels, with positive bootstrap shares equal to one in the reported DR comparisons. WIS/SNIPS results further clarify why IPS can be more variable: predictive policies select users with high weighted control outcomes, making incremental gains difficult to establish under a pure weighting estimator. Weight-trimming checks show

Table 15 Bootstrap uncertainty for policy-value differences

Budget	Metric	Mean	95% CI lower	95% CI upper	Positive bootstrap share
1%	IPS difference	-0	−0.0012	0.001	0.51
1%	DR difference	0.0043	0.0036	0.005	1
5%	IPS difference	−0.0034	−0.0055	−0.0008	0
5%	DR difference	0.0117	0.0103	0.0133	1
10%	IPS difference	−0.0084	−0.0127	−0.005	0
10%	DR difference	0.0155	0.0135	0.0179	1
20%	IPS difference	0.0089	0.0051	0.0136	1
20%	DR difference	0.0227	0.0195	0.0253	1

Table 16 WIS/SNIPS comparison at 10% budget

Policy	IPS	SNIPS difference	Population-scaled WIS	DR	SNIPS treated outcome	SNIPS control outcome
predictive_lgbm	−0.0369	−0.0393	−0.0039	−0.0095	0.7121	0.7514
predictive_rf	−0.011	−0.0248	−0.0025	−0.0033	0.6666	0.6915
uplift_r_linear	−0.0192	0.108	0.0108	0.0123	0.5011	0.393

Table 17 Weight-trimming robustness at 10% budget

Weight setting	Diff. IPS	Diff. DR	IPS positive	DR positive
No trimming	−0.0082	0.0156	False	True
Clip 0.01–0.99	−0.0082	0.0156	False	True
Clip 0.05–0.95	−0.0081	0.0156	False	True
Propensity quantile clip 0.025–0.975	−0.0079	0.0156	False	True

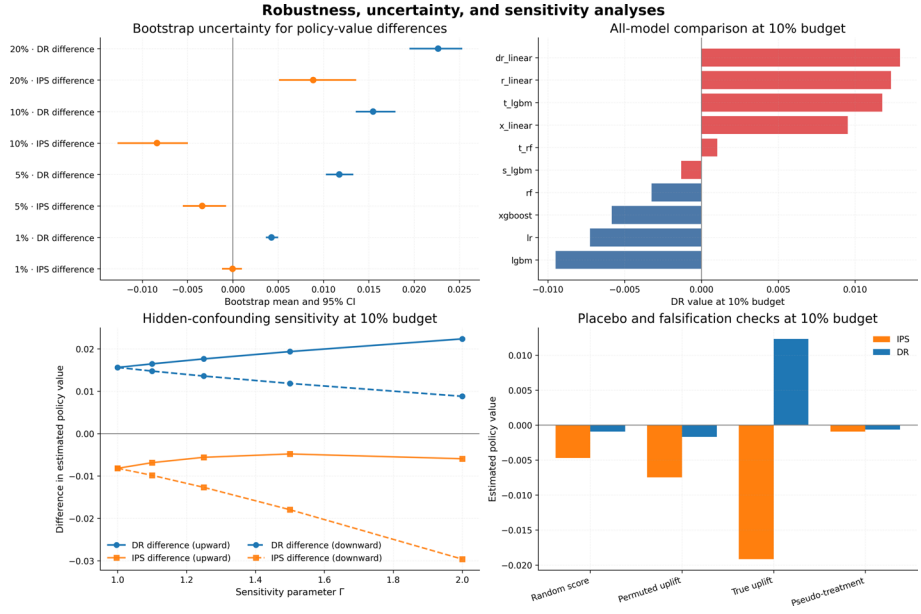

Fig. 8 Bootstrap uncertainty and hidden-confounding sensitivity

Table 18 Hidden-confounding sensitivity at 10% budget

1.00 upward	−0.0082	0.0156	True
1.00 downward	−0.0082	0.0156	True
1.10 upward	−0.0068	0.0164	True
1.10 downward	−0.0099	0.0147	True
1.25 upward	−0.0056	0.0176	True
1.25 downward	−0.0127	0.0135	True
1.50 upward	−0.0048	0.0193	True
1.50 downward	−0.018	0.0118	True
2.00 upward	−0.0059	0.0223	True
2.00 downward	−0.0296	0.0088	True

Table 19 Placebo and falsification tests at 10% budget

Test	IPS value	DR value	Expected pattern
random_score	−0.0047	−0.0009	Random/permuted weaker than true uplift
permuted_uplift_score	−0.0075	−0.0017	Random/permuted weaker than true uplift
true_uplift_score	−0.0192	0.0123	Random/permuted weaker than true uplift
pseudo_treatment_permutation	−0.0009	−0.0006	Pseudo-treatment should not support substantive causal ranking

that the positive DR/AIPW difference remains stable under alternative clipping rules. Together, these checks support the use of DR/AIPW as the main policy-value estimator and IPS as a conservative auxiliary estimator.

4.9 Hidden-confounding sensitivity and placebo tests

Unmeasured confounding cannot be fully ruled out in an observational design. To assess the stability of the main comparison, a hidden-confounding sensitivity analysis perturbs treatment-assignment odds using Γ. Placebo and falsification tests further examine whether the observed uplift advantage can be reproduced by random or mechanically induced rankings.

The hidden-confounding sensitivity analysis is reported in Table 18. Placebo and falsification tests based on random scores, permuted uplift scores, and pseudo-treatment permutations are reported in Table 19.

The DR/AIPW difference remains positive under all evaluated Γ perturbations. Random scores, permuted uplift scores, and pseudo-treatment permutations do not reproduce the policy value of the true uplift score. These results do not eliminate all hidden-confounding concerns, but they reduce the likelihood that the main finding is merely a mechanical artifact of the ranking or evaluation procedure.

4.10 ABCD response-type mechanism diagnostics

Response-type diagnostics help explain why predictive and uplift-based policies select different users under the same budget. These response types are inferred from estimated potential-outcome probabilities and should be interpreted as mechanism diagnostics rather than ground-truth causal labels.

Table 20 reports the ABCD response-type composition of selected users under predictive and uplift policies. Table 21 reports overlap and selection-difference diagnostics, while Figure 9 shows the response-type composition and selection contrast between predictive and uplift-based targeting.

Table 20 Inferred response-type composition of selected users

Policy	Budget	Response type	Share	Selected N
predictive_rf	5%	Persuadable	18.5%	7722
predictive_rf	5%	Sure thing	25.3%	7722
predictive_rf	5%	Lost cause	0.0%	7722
predictive_rf	5%	Do-not-disturb	56.2%	7722
predictive_rf	10%	Persuadable	18.5%	15,445
predictive_rf	10%	Sure thing	26.9%	15,445
predictive_rf	10%	Lost cause	0.0%	15,445
predictive_rf	10%	Do-not-disturb	54.6%	15,445
predictive_rf	20%	Persuadable	28.7%	30,890
predictive_rf	20%	Sure thing	25.4%	30,890
predictive_rf	20%	Lost cause	0.0%	30,890
predictive_rf	20%	Do-not-disturb	45.9%	30,890
uplift_r_linear	5%	Persuadable	49.4%	7722
uplift_r_linear	5%	Sure thing	11.9%	7722
uplift_r_linear	5%	Lost cause	28.4%	7722
uplift_r_linear	5%	Do-not-disturb	10.4%	7722
uplift_r_linear	10%	Persuadable	46.8%	15,445
uplift_r_linear	10%	Sure thing	11.9%	15,445
uplift_r_linear	10%	Lost cause	28.3%	15,445
uplift_r_linear	10%	Do-not-disturb	12.9%	15,445
uplift_r_linear	20%	Persuadable	39.3%	30,890
uplift_r_linear	20%	Sure thing	13.3%	30,890
uplift_r_linear	20%	Lost cause	28.6%	30,890
uplift_r_linear	20%	Do-not-disturb	18.8%	30,890

Table 21 Policy-selection overlap between predictive and uplift policies

Comparison	Budget	Overlap N	Overlap rate	Jaccard	Pred.-only	Uplift-only
Std. LGBM vs uplift	10%	748	4.8%	0.025	0.952	0.952
Std. LGBM vs uplift	20%	3968	12.8%	0.069	0.872	0.872
Strong RF vs uplift	10%	378	2.4%	0.012	0.976	0.976
Strong RF vs uplift	20%	4108	13.3%	0.071	0.867	0.867

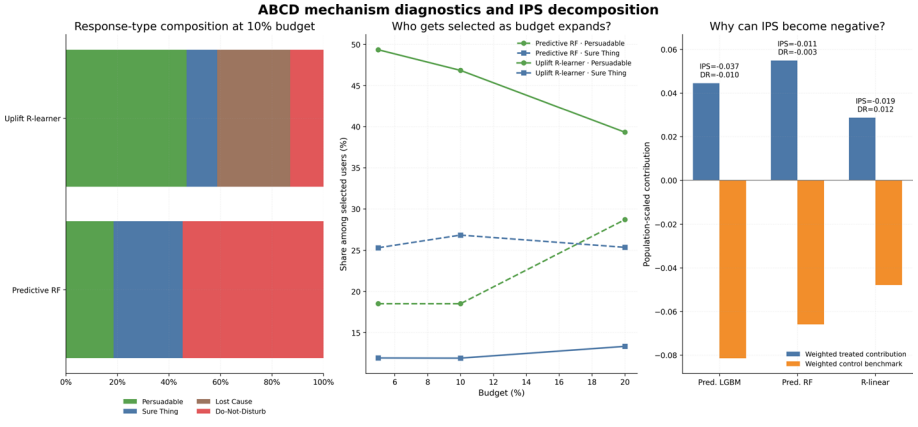

Fig. 9 ABCD composition, policy overlap, and selected-group decomposition

The composition results show that predictive policies select a larger share of users with high baseline response probability, including Sure Things and Do-Not-Disturb users. In contrast, the uplift policy selects a substantially larger share of Persuadables. The low overlap between predictive and uplift selections further indicates that the two scoring objectives allocate the same intervention budget to materially different users. This mechanism evidence helps explain why response-probability ranking and incremental-value ranking lead to different policy values.

4.11 Explaining negative IPS values

Some IPS estimates are weak or negative. This pattern is consistent with the estimator's construction and does not imply that the selected users have low raw response rates. IPS estimates incremental policy value by comparing propensity-weighted treated outcomes with the corresponding weighted control benchmark among selected users. If selected users also have high control outcomes, the resulting IPS value can be small or negative even when raw observed response is high.

Table 22 decomposes selected-group outcomes and weighted control benchmarks to clarify the source of negative or mixed IPS estimates.

The selected-group decomposition shows that predictive policies select users with high observed next-month response rates in both treated and control groups. This indicates that many selected users may borrow even without intervention. By contrast, the uplift policy selects users with lower baseline response but larger estimated incremental value under DR/AIPW. This decomposition helps explain the negative or mixed IPS estimates and supports the interpretation that DR/AIPW, together with WIS/SNIPS and mechanism diagnostics, provides the more relevant basis for policy-value comparison in this observational setting.

4.12 Temporal deployment illustration: cumulative utility under repeated Top-k targeting

To complement the single-period policy-value comparison, we provide an accounting-style temporal deployment illustration under repeated monthly top-k targeting. The purpose of this analysis is not to estimate a dynamic causal effect or to model sequential policy learning. Instead, it asks a narrower operational question: if the same scoring logic were repeatedly applied month by month under a fixed budget rule, how would raw next-month responses and estimated policy value accumulate over time?

Table 22 Selected-group decomposition at 10% budget

Panel A. Selected-group response and weighted benchmarks

Policy	Sel. T share	Raw T	Raw C	SNIPS T	SNIPS C
predictive_lgbm	0.1813	0.7264	0.7537	0.7121	0.7514
predictive_rf	0.2764	0.6554	0.7025	0.6666	0.6915
uplift_r_linear	0.3871	0.4897	0.3513	0.5011	0.3930

Panel B. Incremental policy-value estimates

Policy	IPS	DR/AIPW
predictive_lgbm	−0.0369	−0.0095
predictive_rf	−0.0110	−0.0033
uplift_r_linear	−0.0192	0.0123

Raw T and Raw C denote observed treated and control outcome rates within the selected group. SNIPS T and SNIPS C denote self-normalized weighted treated and control outcome benchmarks. IPS and DR/AIPW are incremental policy-value estimates

Table 23 Final cumulative deployment summary under repeated top-k targeting

Panel A. Observed next-month hits			
Budget	**Pred. hits**	**Uplift hits**	**Diff.**
5%	27,432	13,039	-14,393
10%	53,356	30,414	-22,942
20%	105,059	66,787	-38,272
Panel B. Cumulative DR/AIPW policy value			
Budget	**Pred. DR/AIPW**	**Uplift DR/AIPW**	**Diff.**
5%	-1,380.7	10,746.8	12,127.5
10%	-4,489.5	14,161.1	18,650.6
20%	-9,149.5	17,628.4	26,777.9
Panel C. Cumulative IPS policy value			
Budget	**Pred. IPS**	**Uplift IPS**	**Diff.**
5%	-6,387.5	-7,053.5	−666.0
10%	-18,949.4	-14,624.8	4,324.6
20%	-38,891.5	-28,265.0	10,626.5

Observed hits are descriptive next-month drawdown counts among selected users. DR/AIPW and IPS are cumulative policy-value estimates. DR/AIPW is used as the main policy-value estimator, while IPS is reported as a conservative auxiliary estimator.

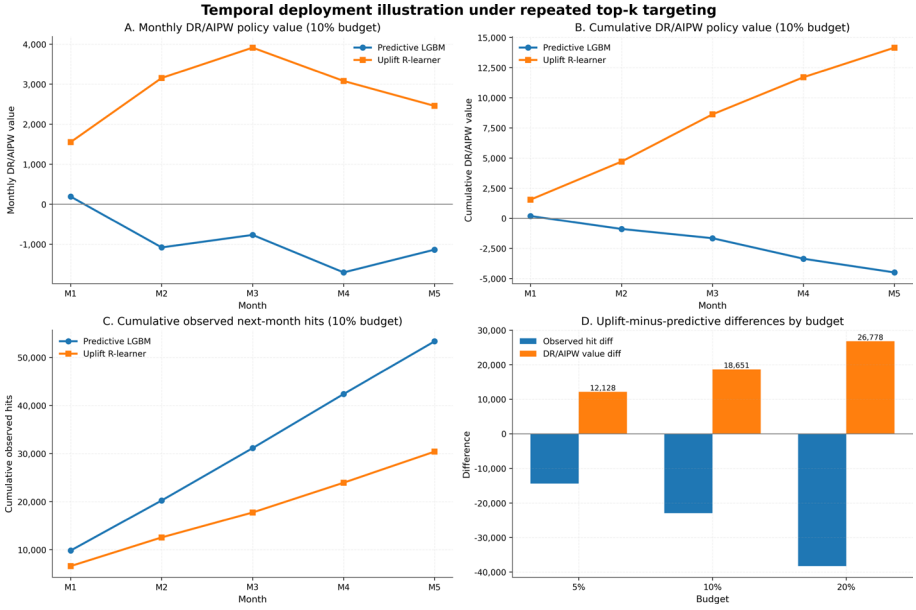

Fig. 10 Cumulative temporal deployment illustration under repeated top-k targeting

In each monthly decision period, users are ranked either by the predictive LightGBM score or by the R-learner uplift score. The top-k users are selected according to the same budget level, and the following month's drawdown outcome is recorded. Two cumulative quantities are then compared. The first is the number of observed next-month hits, which is a descriptive raw response measure. The second is the cumulative DR/AIPW policy value, which estimates the incremental intervention value induced by the corresponding targeting rule.

Table 23 reports the final cumulative summary across the repeated monthly deployment illustration. Figure 10 combines the main cumulative trajectories and shows the contrast between raw response accumulation and DR/AIPW policy value under the two scoring strategies.

The results show that the two scoring strategies optimize different operational quantities. Predictive LightGBM accumulates more observed next-month hits, which is consistent with its objective of ranking users by response probability. However, the R-learner uplift policy accumulates higher DR/AIPW policy value across the evaluated budget levels. This contrast is central to the interpretation of the paper: observed hits indicate how many selected users borrowed in the next month, but they do not distinguish users who borrowed because of the intervention from users who would likely have borrowed anyway.

The temporal deployment illustration therefore reinforces, rather than replaces, the main single-period policy-value evidence. Predictive scoring is better aligned with raw response accumulation, whereas uplift-based scoring is better aligned with estimated incremental intervention value. Since this subsection is descriptive and accounting-style, the results should not be interpreted as evidence about adaptive treatment effects, treatment-state transitions, or policy learning. Its role is to show how the same difference in scoring objectives can translate into different cumulative operational outcomes under repeated budget-constrained targeting.

4.13 Summary of empirical evidence

The empirical results show that predictive scoring and uplift-based scoring should not be evaluated as if they served the same decision objective. The predictive baselines achieve reasonable next-month response prediction performance, with non-trivial AUC, PR-AUC, and calibration results. This confirms that the comparison is not driven by weak predictive benchmarks. Rather, the difference arises because predictive models rank users by response probability, whereas uplift-based models rank users by estimated incremental intervention value.

Under the unified top-k protocol, uplift-based policies achieve higher DR/AIPW policy value than predictive policies across the evaluated budget levels, including comparisons against both the standard LightGBM baseline and the strongest predictive baseline. This pattern provides the main empirical support for the paper's central claim: when the operational objective is budget-constrained intervention allocation, ranking users by incremental treatment value is better aligned with the decision problem than ranking them by raw response probability.

The auxiliary diagnostics are consistent with this interpretation. IPS estimates are more conservative and occasionally weak or negative because they compare selected treated outcomes against a propensity-weighted control benchmark rather than raw response rates. WIS/SNIPS, selected-group decomposition, and ABCD response-type diagnostics clarify this pattern: predictive policies tend to select users with high baseline response probability, including Sure Things and Do-Not-Disturb users, whereas uplift-based policies select a larger share of Persuadables.

The temporal deployment illustration further separates raw response accumulation from incremental intervention value. Predictive scoring accumulates more observed next-month hits, which is expected given its response-probability objective. In contrast, uplift-based scoring accumulates higher DR/AIPW policy value under repeated monthly top-k allocation. This illustration is descriptive and accounting-style rather than an analysis of adaptive or sequential treatment effects, but it reinforces the distinction between observed response and estimated incremental value.

Taken together, the results suggest that uplift-based scoring is better aligned with budget-constrained intervention allocation in this observational internet-lending setting. This conclusion should not be read as a universal superiority claim for uplift models. It is a decision-oriented finding: when the objective is incremental impact under a fixed intervention budget, model evaluation should move beyond predictive accuracy and focus on the policy value induced by the scoring rule.

5 Discussion and implications

The empirical results should be interpreted as evidence about decision alignment under budget-constrained targeting, rather than as a general contest between predictive and causal models. In many operational settings, the relevant question is not only which users are most likely to respond, but which users should receive an intervention when only a limited fraction of the population can be targeted. This distinction is central to the study. By placing predictive and uplift-based scores under the same top-k protocol, the comparison shifts from model accuracy to the policy value induced by each scoring rule.

The findings show that this distinction is empirically meaningful. The predictive models are not weak baselines: their AUC and calibration-related diagnostics indicate reasonable ability to rank next-month response probability. Nevertheless, once these scores are translated into budget-constrained intervention policies, predictive accuracy does not automatically imply higher incremental intervention value. Uplift-based scoring is more directly aligned with the intervention objective because it ranks users by estimated treatment-induced change rather than by response probability. Under the unified top-k protocol, this alignment is reflected in higher DR/AIPW policy value across the evaluated budget levels.

5.1 Reframing targeting as a decision problem

A central implication of the study is that targeting should be framed as a decision problem, not merely as a prediction problem. Predictive models answer the question of who is likely to draw down in the next month. This is useful for forecasting demand, monitoring user behavior, and supporting descriptive or predictive analytics. However, it is not the same question faced by a decision maker who must allocate a limited intervention budget. Under budget constraints, the more relevant question is whose behavior is likely to change because of the intervention.

This reframing explains why a model with acceptable predictive performance may still produce inefficient intervention allocation. A high predicted response probability may reflect a strong baseline tendency to respond, rather than a strong treatment effect. If a user would have drawn down even without intervention, targeting that user may increase observed response counts but contribute little incremental value. Uplift-based scoring addresses a different objective: it attempts to rank users by the difference between treated and untreated potential outcome probabilities. It therefore provides a decision signal that is closer to the operational goal of generating additional responses from a limited intervention budget.

The contribution of this study lies in evaluating this distinction under a controlled decision protocol. All scores are assessed using the same temporal design, budget levels, and top-k rule. As a result, the comparison is not driven by different deployment heuristics, but by the different objectives encoded in predictive and uplift-based scoring.

5.2 Why prediction and intervention allocation diverge

The mechanism diagnostics help explain why predictive and uplift-based policies select different users. Predictive scores tend to prioritize users with high estimated response probability. In settings where many users already have a strong baseline tendency to respond, this can lead predictive targeting to select many Sure Things: users who are likely to respond even without treatment. Such users may improve observed response rates, but they do not necessarily generate high incremental intervention value.

The ABCD diagnostics are useful for interpreting this divergence. Predictive policies select relatively more Sure Things and Do-Not-Disturb users, whereas uplift-based policies select a larger share of Persuadables. These response types should not be interpreted as directly observed causal ground truth. They are inferred from estimated potential-outcome probabilities and are best understood as mechanism-oriented diagnostics. Even with this caution, they clarify the central empirical pattern: predictive and uplift-based scores do not merely reorder the same users; they induce materially different allocation choices.

This divergence becomes especially important under budget constraints. When intervention capacity is large, inefficient allocation may be less visible because many users can be treated. When only a small share of users can be selected, however, spending treatment resources on high-baseline but low-incremental users creates opportunity cost. The value of uplift-based scoring lies in reducing this misalignment between response probability and treatment-induced response.

5.3 Interpreting IPS and DR/AIPW evidence

The mixed and sometimes negative IPS estimates require careful interpretation. In this study, IPS estimates incremental policy value relative to a propensity-weighted control benchmark, not the raw response rate among selected users. A negative IPS estimate does not imply that the selected users have low observed response. It means that, within the selected group, the propensity-weighted treated outcome is lower than the corresponding weighted control benchmark. This can occur when selected users have strong natural response tendencies and the weighted control benchmark is high.

This interpretation is consistent with the selected-group decomposition. Predictive policies tend to select users with high observed response rates in both treated and control groups. As a result, the weighted control benchmark can also be high, making IPS estimates conservative and occasionally negative. The implication is not that the evaluation fails, but that IPS alone is a noisy and conservative basis for drawing conclusions in this observational setting.

For this reason, the study places primary emphasis on DR/AIPW policy value. DR/AIPW combines propensity weighting with outcome regression and is less dependent on weighting alone than IPS. The conclusion therefore does not rely on IPS alone. It is supported by the DR/AIPW advantage of uplift-based policies, while IPS, WIS/SNIPS, bootstrap uncertainty, weight trimming, hidden-confounding sensitivity, placebo tests, and ABCD diagnostics provide complementary evidence.

This evidence chain is stronger than an IPS-centered framing. The results do not require IPS to be positive at every budget level. Instead, the analysis acknowledges the conservativeness of IPS and evaluates whether the uplift-based policy is better aligned with incremental intervention value across multiple diagnostics.

5.4 Practical and methodological implications

For practitioners, the findings suggest that predictive and uplift-based models should not be treated as interchangeable tools. Predictive models remain valuable for forecasting, demand monitoring, risk assessment, and general behavioral analysis. Their limitation appears when response-probability scores are used directly for intervention allocation under a fixed budget. In that setting, the operational objective is not to identify users who are likely to respond, but to identify users whose response probability is most likely to change because of treatment.

A practical decision system should therefore separate forecasting from intervention allocation. Predictive scores can support baseline demand estimation and business monitoring, while uplift scores can guide budget-constrained targeting. The composition of selected users should also be monitored. If a policy repeatedly selects users who would likely respond without intervention, the organization may be spending resources on behavior that would have occurred naturally. If it selects users with weak or negative responsiveness, treatment may be inefficient or even counterproductive.

Methodologically, the study reinforces the importance of evaluating scoring models at the decision level. AUC, calibration, raw AUUC, and Qini provide useful diagnostic information, but none of them alone answers the deployment question. The relevant object is the policy induced by a score once it is embedded into a budget-constrained top-k rule. Holding the decision rule fixed allows the comparison to focus on whether the scoring objective is aligned with the intervention objective.

The results also illustrate why raw uplift-ranking metrics should be treated cautiously in observational data. AUUC and Qini may indicate whether a score contains useful uplift-ranking signal, but they are not sufficient as standalone causal decision criteria when treatment assignment is non-random. In such settings, adjusted policy-value estimators and robustness diagnostics are necessary. The additional nonlinear R- and DR-learner checks suggest that more flexible treatment-effect models may further improve performance, but these checks should be interpreted as model-capacity evidence unless the full pipeline is rerun with those learners pre-specified.

5.5 Limitations and future research

Several limitations should be noted. The analysis is based on a single observational internet-lending dataset, so the findings should be interpreted as dataset-specific empirical evidence rather than as a universal claim that uplift-based scoring always outperforms predictive scoring. Although the sample is large and temporally structured, external validation on additional datasets would be necessary to assess generalizability.

Causal interpretation depends on standard assumptions, including consistency, conditional exchangeability, and positivity. The study uses pre-treatment covariates, propensity diagnostics, covariate-balance checks, DR/AIPW estimation, sensitivity analysis, and placebo tests to improve credibility. These steps reduce some concerns, but they cannot eliminate the possibility of unmeasured confounding.

The outcome is defined as next-month drawdown to improve temporal ordering and reduce same-period leakage. This design is stricter than evaluating contemporaneous outcomes, but it does not capture all longer-term consequences of intervention, such as repeated borrowing, credit risk, profitability, user fatigue, or customer satisfaction.

Similarly, the treatment is represented as a binary current-month intervention indicator, whereas real intervention systems often vary by channel, timing, amount, and content.

Future research can extend the framework in several directions. Multi-treatment settings would allow policies to choose not only whom to target, but also which intervention to deliver. Sequential and dynamic settings would allow researchers to examine how past interventions affect future responsiveness and user behavior. Another promising direction is to pre-specify nonlinear treatment-effect learners as formal candidates and evaluate them through the same train-validation-evaluation protocol. These extensions would help determine whether the decision-oriented conclusions observed here persist in richer and more dynamic intervention environments.

6 Conclusion

This study examined budget-constrained targeting as a decision problem. Rather than proposing a new causal estimator, it compared predictive and uplift-based scoring under the same top-k decision protocol, where limited intervention capacity requires users to be ranked for treatment.

The results show that predictive accuracy and intervention value can diverge. The predictive models are not weak baselines; they achieve reasonable next-month response prediction. However, response probability is not equivalent to incremental intervention value. Under the unified budget-constrained protocol, uplift-based policies achieve higher DR/AIPW policy value than predictive policies across the evaluated budget levels, including comparisons against the strongest predictive baseline.

The mixed or negative IPS estimates do not overturn this conclusion. In this setting, IPS is a conservative auxiliary estimator because selected predictive-policy users often have high weighted control benchmarks. The broader evidence from DR/AIPW policy value, WIS/SNIPS, bootstrap uncertainty, weight trimming, hidden-confounding sensitivity, placebo tests, and ABCD mechanism diagnostics supports the interpretation that uplift-based scoring is better aligned with incremental intervention allocation in the studied observational setting.

The contribution of the study is empirical and decision-oriented. It shows that when model scores are embedded into the same budget-constrained decision rule, the scoring objective can materially change allocation outcomes. The findings should not be read as evidence that uplift models are universally superior. Rather, they show that when the operational objective is incremental impact under a fixed intervention budget, evaluation should move beyond predictive accuracy and focus on the policy value induced by the scoring rule.

Table 24 Observed next-month hit accumulation under repeated monthly top-(k) targeting

Budget	Month	Pred. Hits	Uplift Hits	Diff	Cum. Pred. Hits	Cum. Uplift Hits	Cum. Diff
5%	M1	5028	3298	−1730	5028	3298	−1730
5%	M2	5370	2598	−2772	10,398	5896	−4502
5%	M3	5604	2153	−3451	16,002	8049	−7953
5%	M4	5794	2415	−3379	21,796	10,464	−11,332
5%	M5	5636	2,575	−3061	27,432	13,039	−14,393
10%	M1	9835	6605	−3230	9835	6605	−3230
10%	M2	10,430	5966	−4464	20,265	12,571	−7694
10%	M3	10,877	5194	−5683	31,142	17,765	−13,377
10%	M4	11,244	6192	−5052	42,386	23,957	−18,429
10%	M5	10,970	6457	−4513	53,356	30,414	−22,942
20%	M1	19,081	12,797	−6284	19,081	12,797	−6284
20%	M2	20,332	13,120	−7212	39,413	25,917	−13,496
20%	M3	21,173	12,695	−8478	60,586	38,612	−21,974
20%	M4	22,301	14,077	−8224	82,887	52,689	−30,198
20%	M5	22,172	14,098	−8074	105,059	66,787	−38,272

Table 25 DR/AIPW cumulative policy-value decomposition under repeated monthly top-(k) targeting

Budget	Month	Pred. DR/AIPW	Uplift DR/AIPW	Diff	Cum. Pred. DR/AIPW	Cum. Uplift DR/AIPW	Cum. Diff
5%	M1	348.8	972.0	623.1	348.8	972.0	623.1
5%	M2	−278.9	2102.5	2381.4	69.9	3074.5	3004.6
5%	M3	−110.4	2777.9	2888.4	−40.6	5852.4	5893.0
5%	M4	−787.1	2574.2	3361.3	−827.7	8426.6	9254.2
5%	M5	−553.0	2320.3	2873.3	−1380.7	10,746.8	12,127.5
10%	M1	193.3	1553.1	1359.8	193.3	1553.1	1359.8
10%	M2	−1078.7	3155.3	4234.0	−885.4	4708.5	5593.9
10%	M3	−766.5	3914.6	4681.1	−1651.9	8623.1	10,274.9
10%	M4	−1703.3	3,080.3	4783.7	−3355.2	11,703.4	15,058.6
10%	M5	−1134.4	2,457.7	3592.1	−4489.5	14,161.1	18,650.6
20%	M1	−332.5	3047.5	3380.0	−332.5	3047.5	3380.0
20%	M2	−1606.4	3972.7	5579.1	−1938.9	7020.2	8959.1
20%	M3	−1108.2	4448.4	5556.6	−3047.1	11,468.6	14,515.7
20%	M4	−3509.7	3376.6	6886.3	−6556.8	14,845.2	21,402.0
20%	M5	−2592.7	283.2	5376.0	−9149.5	17,628.4	26,777.9

Appendix A. Monthly decomposition for the accounting-style temporal deployment illustration

This appendix provides the month-by-month decomposition supporting the accounting-style temporal deployment illustration reported in Sect. 4.12. These results are intended to improve transparency for the repeated monthly top-k targeting analysis. They should not be interpreted as evidence about adaptive treatment effects, sequential policy learning, or treatment-state transitions.

Appendix Table 24 reports the accumulation of observed next-month hits under predictive LightGBM scoring and R-learner uplift scoring. Observed hits are descriptive raw responses among selected users. Appendix Table 25 reports the corresponding DR/AIPW policy value, which estimates the incremental intervention value induced by each targeting rule. The main-text interpretation remains based on the final cumulative summary in Table 23 and the combined visualization in Fig. 10.

Author contributions
J.J. conceived the study, designed the empirical framework, conducted the data analysis, and drafted the manuscript. N.A.W.A.H. contributed to the methodological design, interpretation of results, and critical revision of the manuscript. N.K.Y. assisted with model implementation, experimental validation, and manuscript editing. W.C.C. supervised the research process, provided conceptual guidance, and contributed to manuscript revision. All authors reviewed and approved the final manuscript.

Funding
This research did not receive any specific grant from public, commercial, or not-for-profit funding agencies.

Data availability
The data used in this study are subject to confidentiality and commercial restrictions and are therefore not publicly available. Aggregated results and derived analysis outputs are reported within the manuscript. Further information may be provided by the corresponding author upon reasonable request, subject to data-sharing approval.

Code availability
The complete source code cannot be made publicly available because it is linked to confidential data-processing procedures, internal variable structures, and data-provider restrictions. Non-sensitive code excerpts, model specifications, and evaluation procedures may be made available from the corresponding author upon reasonable request, subject to confidentiality obligations, data-provider requirements, and relevant institutional approval.

Declarations

Ethics approval and consent to participate
This study did not involve prospective human-subject experiments, recruitment of human participants, collection of biological materials or human tissue samples, or direct administration of interventions by the authors. The study used retrospective, de-identified secondary operational data provided by Haier Consumer Finance Co., Ltd. for academic research purposes. The data were generated from routine platform operations before the research analysis was conducted. Before being provided to the authors, the data were de-identified by the data provider. The authors did not access names, identity numbers, telephone numbers, addresses, bank account information, or other directly identifiable personal information. All analyses were conducted on de-identified user-month records, and all results are reported only in aggregated or derived analytical form. The study was conducted in accordance with applicable institutional requirements, data-provider restrictions, confidentiality obligations, and relevant regulations governing the use of de-identified secondary operational data.

Accordance
The authors confirm that all methods were carried out in accordance with relevant institutional, legal, regulatory, data-provider, and confidentiality requirements for research using de-identified secondary operational data. The analysis was conducted using de-identified data only, and no directly identifiable personal information was accessed, reported, or disclosed.

Consent for publication
Not applicable. The manuscript does not contain personally identifiable information, individual-level personal data, images, clinical case materials, or any other materials requiring consent for publication. All results are presented in aggregated or derived analytical form.

Informed consent
Not applicable to the authors' secondary analysis. This study was based on retrospective, de-identified secondary operational records. The authors did not directly recruit, contact, or interact with individual participants, and no new personal data were collected by the authors for this study.

Competing interests
The authors declare no conflict of interest.

Received: 1 February 2026 / Accepted: 4 June 2026
Published online: 14 June 2026

References

1. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc. 2012;107:1106–18. https://doi.org/10.1080/01621459.2012.695674.
2. Kallus N. Balanced policy evaluation and learning. In: Advances in neural information processing systems. 2018. p. 31.
3. Bertsimas D, Kallus N. From predictive to prescriptive analytics. Manage Sci. 2019;66:1025–44. https://doi.org/10.1287/mnsc.2018.3253.
4. Athey S, Wager S. Policy learning with observational data. Econometrica. 2021;89:133–61. https://doi.org/10.3982/ECTA15781.
5. Rich E. User modeling via stereotypes. Cogn Sci. 1979;3:329–54. https://doi.org/10.1207/s15516709cog0304_3.
6. Brusilovsky P. Adaptive hypermedia. User Model User-Adap Inter. 2001;11:87–110. https://doi.org/10.1023/A:1011143116306.
7. Ricci F, Rokach L, Shapira B. Recommender Systems Handbook. Springer Boston. 2011. https://doi.org/10.1007/978-0-387-85820-3.
8. Verbeke W, Martens D, Baesens B. Social network analysis for customer churn prediction. Decis Support Syst. 2014;70:431–46. https://doi.org/10.1016/j.dss.2014.05.007.

9. Eke CI, Norman AA, Shuib L, Nweke HF. A survey of user profiling: State-of-the-art, challenges, and solutions. IEEE Access. 2019;7:144907–24. https://doi.org/10.1109/ACCESS.2019.2947003.
10. Purificato E, Boratto L, Luca D, EW. User modeling and user profiling: a comprehensive survey. arXiv. 2024.arXiv:2402.09660.
11. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688–701. https://doi.org/10.1037/h0037350.
12. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55. https://doi.org/10.1093/biomet/70.1.41.
13. Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge: Cambridge University Press; 2015. https://doi.org/10.1017/CBO9781139025751.
14. Lo VSY. The true lift model: a novel data mining approach to response modeling in database marketing. ACM SIGKDD Explor Newsl. 2002;4:78–86. https://doi.org/10.1145/772862.772873.
15. Radcliffe NJ, Surry PD. Real-world uplift modelling with significance-based uplift trees. Stochastic Solutions White Paper; 2011.
16. Gutiérrez P, Gérardy J-Y. Causal inference and uplift modelling: a review of the literature. In: Proceedings of the international conference on predictive applications and APIs. PMLR. 2017. p. 1–13.
17. Devriendt F, Moldovan D, Verbeke W. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling. Big Data. 2018;6:13–41. https://doi.org/10.1089/big.2017.0074.
18. Rzepakowski P, Jaroszewicz S. Decision trees for uplift modeling with single and multiple treatments. Knowl Inf Syst. 2012;32:303–27. https://doi.org/10.1007/s10115-011-0434-0.
19. Guelman L, Guillén M, Pérez-Marín AM. Uplift modeling for churn prevention in telecommunications. Decis Support Syst. 2015;74:10–9. https://doi.org/10.1016/j.dss.2015.04.002.
20. Michel A, Sakarovitch B, Valette F. Tree-based uplift modeling for marketing. Expert Syst Appl. 2017;70:351–64. https://doi.org/10.1016/j.eswa.2016.11.018.
21. Olaya D, Coussement K, Verbeke W. A survey and benchmarking study of multitreatment uplift modeling. Data Min Knowl Disc. 2020;34:273–308. https://doi.org/10.1007/s10618-019-00670-8.
22. Athey S, Imbens GW. Recursive partitioning for heterogeneous causal effects. Proc Nat Acad Sci USA. 2016;113:7353–60. https://doi.org/10.1073/pnas.1510489113.
23. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc. 2018;113:1228–42. https://doi.org/10.1080/01621459.2017.1319839.
24. Athey S, Tibshirani J, Wager S. Generalized random forests. Ann Stat. 2019;47:1148–78. https://doi.org/10.1214/18-AOS1709.
25. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci USA. 2019;116:4156–65. https://doi.org/10.1073/pnas.1804597116.
26. Shalit U, Johansson FD, Sontag D. Estimating individual treatment effect: Generalization bounds and algorithms. In: Proceedings of the 34th international conference on machine learning (ICML). PMLR. 2017. p. 3076–85.
27. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Econom J. 2018;21:C1–68. https://doi.org/10.1111/ectj.12097.
28. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika. 2021;108:299–319. https://doi.org/10.1093/biomet/asaa076.
29. Sverdrup E, Cui Y. Proximal causal learning of conditional average treatment effects. In: Proceedings of the international conference on machine learning. PMLR. 2023. p. 33285–98.
30. Rafieian O, Yoganarasimhan H. AI and personalization. Mark Sci. 2023;42:77–102. https://doi.org/10.1287/mksc.2022.1383.
31. Jiang J, Hamid MNAWA, Yap NK, Chong CW. A dynamic framework for causal user profiling and treatment segmentation via uplift modeling in internet lending. IEEE Access. 2026;14:40147–71. https://doi.org/10.1109/ACCESS.2026.3670857.

Publisher's Note