Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Uncertainty-Aware Uplift Modeling for Safer Marketing Targeting: Conformal Prediction and Bayesian Calibration with LCB Policies

Yifei Lu[1], *Jinyi Mu*[2], *Thao Tran*[3]
[1]*Computer Science, UCSD, CA, USA*
[2]*Computer Science and Engineering, UCSD, CA, USA*
[3]*Data Science, University of Pittsburgh, PA, USA*
yifei.lu0816@outlook.com

DOI: 10.69987/JACS.2024.40507

Keywords

uplift modeling;
conditional average treatment effect;
conformal prediction;
Bayesian calibration;
lower confidence bound;
off-policy evaluation;
marketing targeting

Abstract

Uplift models estimate the incremental impact of an intervention (e.g., a marketing message) on an outcome, and are widely used to allocate limited treatment budget. In practice, incremental effects are small and noisy, so point-estimate uplift targeting can be unstable: aggressive targeting may "flip" ROI from positive to negative when the true uplift is near zero. This paper studies uncertainty-aware uplift targeting for operational risk control. We convert point uplift estimates into calibrated uncertainty intervals using two complementary approaches: (i) normalized split conformal prediction on transformed-outcome regression, yielding distribution-free marginal coverage for pseudo-outcomes; and (ii) Bayesian binned calibration, yielding interpretable credible intervals for bin-level uplift via Beta posteriors over treated/control conversion rates. We then deploy a conservative lower confidence bound (LCB) policy that ranks or filters customers by LCB uplift to trade expected gain for reduced downside risk. We conduct full empirical evaluations on the Kevin Hillstrom MineThatData e-mail campaign dataset (64,000 customers) under two binary-treatment tasks: Mens E-Mail vs No E-Mail and Womens E-Mail vs No E-Mail, with conversion as the primary business outcome. For 90% nominal intervals, normalized conformal achieved 0.904 and 0.899 marginal coverage on the transformed outcome for the Mens and Womens tasks, respectively, while producing heterogeneous interval widths that reflect feature-dependent noise. In policy evaluation via inverse propensity scoring, LCB-conformal targeting improved the best 95% lower bound profit from 15.1 to 21.9 per 10,000 customers on the Mens task compared with point-estimate targeting at the same cost assumptions. On the Womens task, point-estimate targeting delivered the highest risk-adjusted profit, while LCB policies were more conservative and reduced expected profit. Overall, uncertainty estimates are essential for explaining and controlling risk; when interval width correlates with ranking mistakes, LCB targeting materially reduces "bad surprise" outcomes.

Introduction

Marketing and advertising systems increasingly rely on individualized decision rules to select who should receive a treatment—an e-mail offer, a discount, an advertisement impression, or a push notification. In randomized controlled experiments, the relevant quantity is not the response probability under treatment alone, but the *incremental* effect of treatment relative to control. Uplift modeling (also called differential response modeling) operationalizes this idea by learning conditional average treatment effects (CATE) from A/B test logs and ranking individuals by predicted uplift [1]–[3].

Despite its conceptual appeal, uplift targeting is operationally risky. Incremental effects in mature marketing channels are often small, heterogeneous, and noisy; moreover, the decision is typically made under a fixed budget or a hard cost-per-contact constraint. When the true uplift is close to zero, ranking errors can cause

"over-targeting": customers with negative or near-zero true uplift are treated, driving down ROI and potentially increasing churn or fatigue. In production systems, this manifests as unstable campaign performance across time, geographies, or creatives—precisely the "flip" outcomes that practitioners describe as campaign 'crashes'.

Risk control requires moving beyond point estimates. Standard supervised learners can output a point prediction of uplift, but they do not quantify uncertainty in a way that is directly usable for decision-making. Two additional problems occur in practice. First, uplift models can be poorly calibrated: the predicted uplift scale may not match realized incremental lift, especially when the model is tuned for ranking rather than absolute effect size [21]–[23]. Second, even when mean predictions are accurate on average, the uncertainty can vary substantially across the feature space due to differing sample sizes, covariate overlap, and outcome rarity [25-30].

This paper studies an uncertainty-aware uplift pipeline designed specifically for *risk-stable targeting*. Our goal is to attach a principled interval to each uplift prediction and then use a conservative lower confidence bound (LCB) for targeting decisions. The LCB idea is natural: if uplift has an interval estimate [LCB(x), UCB(x)], then treating only when LCB(x) exceeds a cost threshold provides a distribution-free or probabilistic safety margin. Similar lower-bound decision rules appear in bandit optimization and conservative policy improvement, where exploration or deployment is constrained by worst-case performance [12], [13].

We combine two complementary interval construction strategies. First, we use split conformal prediction, a distribution-free framework that guarantees marginal coverage without parametric assumptions [15]–[17]. Because individual causal effects are not directly observed, we apply conformal inference to a transformed outcome whose conditional expectation equals the CATE in randomized experiments. To avoid the overly conservative constant-width intervals of basic split conformal, we use a normalized (locally adaptive) nonconformity score that scales residuals by a learned conditional dispersion model, yielding feature-dependent interval widths [18]. Second, we use Bayesian binned calibration: we bucket individuals by point-estimate uplift score, then estimate treated/control conversion rates in each bucket with Beta posteriors and propagate uncertainty to uplift via Monte Carlo. This produces interpretable credible intervals on bucket-level uplift and acts as a pragmatic calibration layer for ranking outputs [31-36].

To connect uncertainty estimates to campaign operations, we evaluate a family of LCB targeting policies. Under a budget fraction ρ, the policy treats the top-ρ customers ranked by LCB uplift. Under a unit cost c and conversion value v, the policy can also be thresholded as treat-if-LCB(x) > c/v. We evaluate policies using inverse propensity scoring (IPS) on randomized logs, reporting both expected profit and statistical risk (standard error and 95% lower bounds) [11].

Although our motivation is large-scale uplift in advertising systems, the experiments in this paper are conducted on the Kevin Hillstrom MineThatData e-mail campaign dataset, a public randomized marketing benchmark with rich customer features and three campaign segments (Mens E-Mail, Womens E-Mail, and No E-Mail) [4], [5]. In our evaluation environment, the widely used Criteo uplift benchmark is distributed primarily in compressed binary formats (gzip/parquet) that are unavailable for download; therefore we use Hillstrom as a fully accessible randomized dataset while keeping the methodology unchanged. The Hillstrom benchmark is directly relevant to uplift targeting because it contains randomized treatment assignment, a sparse conversion outcome, and realistic covariates such as purchase history and channel [37-42].

Our contributions are as follows:

(1) We present an end-to-end uncertainty-aware uplift workflow that produces point uplift estimates, calibrated intervals, and LCB scores from a single A/B log.
(2) We instantiate two interval methods—normalized conformal prediction on transformed-outcome regression and Bayesian binned calibration on T-learner scores—and quantify their coverage and width on held-out data.
(3) We evaluate LCB targeting policies with IPS-based profit estimation and show how uncertainty intervals translate into practical risk–return curves.
(4) We provide reproducible empirical findings (tables, figures, hyperparameters, and software versions) to support rigorous comparison between point targeting and LCB targeting.

From a causal-inference perspective, uplift corresponds to the conditional difference between two potential outcomes, Y(1) and Y(0), under treatment and control [6]–[10]. Even in randomized experiments, $\tau(x)$ can be difficult to estimate accurately because the observed outcomes are sparse (conversions are rare) and the feature space is high-dimensional. Moreover, the ranking objective is asymmetric: a small number of false positives (customers treated despite negative τ) can erase the gains from many true positives, especially when treatment cost is nontrivial. This asymmetry motivates risk-sensitive decision rules that explicitly consider uncertainty rather than relying solely on mean predictions.

Uncertainty quantification for uplift has two distinct roles. The first is *statistical reporting*: analysts want confidence intervals for incremental lift in segments to communicate whether a campaign is effective. The second is *decision-making*: the production system needs a score that trades off expected gain against the probability of loss. The two roles are related but not identical. A method can provide well-calibrated intervals for reporting yet still be suboptimal for targeting if the interval width is uncorrelated with ranking errors. Conversely, a heuristic uncertainty proxy can improve targeting even if it is not a formal confidence interval.

We therefore focus on uncertainty methods that are both principled and operational. Conformal prediction provides finite-sample guarantees with minimal assumptions [15]–[17], making it attractive when the data-generating process is complex or when the model class is misspecified. Bayesian approaches are attractive when interpretability matters, because the posterior directly quantifies probability mass over plausible uplift values and can be aggregated to business-friendly deciles. In modern marketing stacks, these two approaches are often used separately—conformal in ML uncertainty and Bayesian in analytics. Our contribution is to evaluate both in a unified uplift targeting setting and to connect them to a simple LCB policy that aligns with how campaign managers reason about "safe" incremental lift.

Related work in uplift modeling has explored a variety of learners and loss functions, including uplift trees and forests [2], two-model and single-model (transformed-outcome) approaches, and flexible meta-learners for heterogeneous treatment effect estimation [14]. Large-scale benchmarks such as Criteo's uplift dataset have accelerated empirical comparisons in ad-tech contexts [3]. In contrast, the uncertainty and calibration aspects of uplift are less standardized: many deployments still rely on point uplift rankings with ad-hoc guardrails (e.g., minimum predicted uplift thresholds) that are not tied to statistical coverage. By providing detailed empirical results for interval quality and for policy risk–return curves, this paper aims to make uncertainty a first-class evaluation axis for uplift targeting.

Method

A. Uplift and decision objective.

We observe i.i.d. samples (X i, W i, Y i) from a randomized marketing experiment, where X_i denotes customer features, W i ∈ {0,1} indicates whether the customer received the treatment (campaign contact), and Y_i ∈ {0,1} is the conversion outcome. The conditional average treatment effect (CATE) or uplift is defined as $\tau(x) = E[Y|W=1,X=x] - E[Y|W=0,X=x]$. Given a cost c per treated customer and a value v per conversion, a deterministic targeting policy $\pi(x) \in \{0,1\}$ yields expected net profit $E[v \cdot Y(\pi) - c \cdot \pi(X)]$. In budgeted campaigns, π is constrained to treat a fixed fraction ρ of customers, and the operational problem becomes ranking customers by a score s(x) and treating the top-ρ fraction.

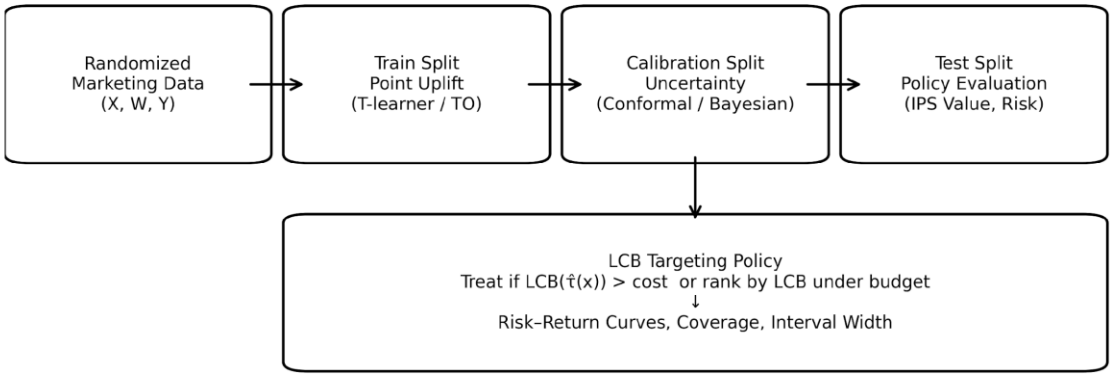

Fig. 1. Point uplift estimation, interval calibration, and LCB targeting with IPS evaluation.

B. Dataset and binary-treatment tasks.

We use the Kevin Hillstrom MineThatData e-mail marketing dataset [4], [5]. It contains 64,000 customers with randomized assignment to one of three segments: Mens E-Mail, Womens E-Mail, or No E-Mail. The dataset includes behavioral covariates (recency, purchase history), demographic indicators (mens, womens, newbie), and categorical attributes (zip_code, channel, history_segment). Outcomes include visit (site visit), conversion (purchase), and spend (purchase amount). Because uplift is defined relative to a single control, we construct two binary-treatment tasks by filtering the dataset:

• Mens task: W=1 for Mens E-Mail and W=0 for No E-Mail (n=42,666).

• Womens task: W=1 for Womens E-Mail and W=0 for No E-Mail (n=42,667).

Throughout, conversion is the primary business outcome Y; visit is treated as an auxiliary metric and spend is used only for descriptive statistics.

Table I. Hillstrom dataset segment-level summary (conversion is the primary outcome).

segment	n	visit_rate	conversion_rate	mean_spend
Mens E-Mail	21307	0.1828	0.0125	1.4226
No E-Mail	21306	0.1062	0.0057	0.6528
Womens E-Mail	21387	0.1514	0.0088	1.0772

C. Preprocessing and data splits.

All models share the same preprocessing pipeline: numeric features are standardized with StandardScaler (without mean centering to support sparse matrices) and categorical features are one-hot encoded (Table II). For each task, we perform a 60/20/20 split into training, calibration, and test sets, stratified by the joint label (W,Y) to preserve both treatment balance and conversion rarity.

Table II. Feature types and preprocessing used in all models.

feature	type	unique_values	preprocessing
recency	numeric	12	StandardScaler (no mean centering)
history	numeric	34833	StandardScaler (no mean centering)
mens	numeric	2	StandardScaler (no mean centering)
womens	numeric	2	StandardScaler (no mean centering)
newbie	numeric	2	StandardScaler (no mean centering)
zip_code	categorical	3	OneHotEncoder (handle_unknown=ignore)
channel	categorical	3	OneHotEncoder (handle unknown=ignore)
history_segment	categorical	7	OneHotEncoder (handle_unknown=ignore)

Table III. Train/calibration/test splits for both tasks (stratified by (W,Y)).

task	split	n	treat_rate	conversion_rate
Mens	Train	25567	0.5000	0.0091
Mens	Calibration	8523	0.5001	0.0092
Mens	Test	8523	0.4999	0.0092
Womens	Train	25615	0.5010	0.0073
Womens	Calibration	8539	0.5010	0.0073
Womens	Test	8539	0.5009	0.0074

D. Point uplift estimation.

We evaluate two standard uplift estimators that are compatible with large-scale marketing logs.
1) T-learner with logistic regression: we fit separate probabilistic outcome models for treated and control groups, $\hat{p}1(x)=P(Y=1|X=x,W=1)$ and $\hat{p}0(x)=P(Y=1|X=x,W=0)$, using L2-regularized logistic regression. The point uplift estimate is $\hat{\tau}\ T(x)=\hat{p}1(x)-\hat{p}0(x)$. This approach directly models conversion probabilities and supports standard probability calibration methods.
2) Transformed-outcome regression (TO-Ridge): for randomized experiments with propensity $e=P(W=1)$, the transformed outcome $Y\sim = Y\cdot(W/e - (1-W)/(1-e))$ satisfies $E[Y\sim|X=x]=\tau(x)$ [6], [7]. We fit a single Ridge regression model $\hat{\mu}(x)=E[Y\sim|X=x]$, and use $\hat{\tau}\_TO(x)=\hat{\mu}(x)$ as the point uplift estimate. TO regression reduces the two-model estimation problem to standard regression and is attractive when the outcome is rare.

In both tasks, treatment assignment is approximately balanced (treat_rate≈0.5 in each split; Table III). This simplifies estimation because the variance of inverse-propensity weights is low. Nevertheless, conversion is rare (≈0.8–1.1%), which makes model selection sensitive to regularization and calibration. We therefore prefer convex models with explicit regularization (logistic regression and Ridge) to avoid overfitting and to keep uncertainty estimates stable. The T-learner is trained with a shared preprocessing map (one-hot categories fit on the full training split) to ensure consistent feature encoding across the treated and control outcome models.

E. Conformal uncertainty intervals on transformed outcomes.

Conformal prediction provides distribution-free predictive intervals with finite-sample marginal coverage under exchangeability [15]–[17]. We apply split conformal on the transformed outcome $Y\sim$, which is observed for each unit and whose conditional expectation equals uplift under randomized assignment. Let $\hat{\mu}$ be the TO-Ridge predictor trained on the training split. On the calibration split, we compute absolute residuals $r\ i=|Y\sim\ i - \hat{\mu}(X\ i)|$. Basic split conformal uses the $(1-\alpha)$-quantile q of $\{r\ i\}$ and outputs $[\hat{\mu}(x)-q, \hat{\mu}(x)+q]$; this yields constant-width intervals that can be overly conservative.

To obtain heterogeneous widths, we use a normalized nonconformity score. We fit an auxiliary dispersion model $\hat{\sigma}(x)$ on the calibration data by regressing $\log(r\_i+\varepsilon)$ on X_i ($\varepsilon=10^{-6}$). We then define scores $s\_i = r\ i / \hat{\sigma}(X\ i)$ and let q be the $(1-\alpha)$-quantile of $\{s\ i\}$. For a new x, the interval is $[LCB\_C(x), UCB\_C(x)] = [\hat{\mu}(x) - q\cdot\hat{\sigma}(x), \hat{\mu}(x) + q\cdot\hat{\sigma}(x)]$.
When $\hat{\sigma}$ is informative, this normalized conformal interval approximates conditional heteroscedasticity while maintaining strong empirical marginal coverage on $Y\sim$ [18]. We use $\alpha=0.1$ (90% nominal) unless stated otherwise.

F. Bayesian binned calibration intervals on T-learner scores.

As a complementary approach, we compute Bayesian credible intervals for bin-level uplift. We first compute the point uplift score $\hat{\tau}\ T(x)$ on the calibration split and partition the calibration customers into $K=10$ equal-frequency bins (deciles) by $\hat{\tau}\ T$. Within each bin b, let $(n1\ b, c1\ b)$ and $(n0\ b, c0\ b)$ denote the treated/control counts and conversions. We place independent Beta(1,1) priors over the treated and control conversion probabilities $p1\ b$ and $p0\ b$, yielding posteriors $p1\ b|D \sim Beta(1+c1\_b,\ 1+n1\_b-c1\_b)$ and $p0\_b|D \sim Beta(1+c0\_b,\ 1+n0\_b-c0\_b)$. The bin uplift $\tau\ b=p1\ b-p0\ b$ is then a derived random variable. We draw 20,000 Monte Carlo samples from the two Beta

posteriors and estimate the 5th/95th percentiles to obtain a 90% credible interval [LCB B(b), UCB B(b)]. For a test customer x, we assign it to a bin b(x) by its $\hat{\tau}$_T(x) score and set [LCB B(x), UCB B(x)] = [LCB B(b(x)), UCB B(b(x))].

This discretized approach is intentionally simple: it provides interpretable uncertainty at the resolution at which campaign analytics are typically reported (deciles or ventiles), and it acts as a calibration layer that links score ranges to realized incremental effects [21]–[23].

G. LCB targeting policies and IPS evaluation.

We compare three targeting scores: (i) point uplift from T-learner logistic regression (Point(T-LR)); (ii) conformal LCB from TO-Ridge (LCB-Conformal); and (iii) Bayesian bin LCB from T-learner bins (LCB-BayesianBin). Given a budget fraction ρ, a policy selects the top-ρ customers ranked by the chosen score and sets π_i=1 for those customers and π_i=0 otherwise.

Because the data arise from randomized treatment assignment, we evaluate each deterministic policy with inverse propensity scoring (IPS) [11]. Let e be the empirical propensity on the training split (≈0.5 in both tasks). For a policy π, the IPS estimate of the expected conversion rate is $\hat{Y}(\pi) = (1/n)\ \Sigma_i\ [\ \pi_i \cdot W_i \cdot Y_i/e + (1-\pi_i)\cdot(1-W_i)\cdot Y_i/(1-e)\ ]$.

We report incremental conversions relative to treating nobody (π≡0) and net profit per customer as $v\cdot(\hat{Y}(\pi)-\hat{Y}(0)) - c\cdot E[\pi]$. Standard errors are computed from the empirical variance of the per-sample IPS contributions. Unless stated otherwise, we set v=1 and c=0.001 (a 0.1-cent cost per contacted customer) to convert incremental conversions into a comparable profit scale.

H. Evaluation metrics.

Point-estimate uplift is primarily evaluated as a *ranking* problem. We compute Qini curves, a standard uplift ranking diagnostic [1], [2]. Given a score s(x), we sort the test set in descending order and consider the prefix of the top-k customers. Let C_t(k) and N_t(k) be the cumulative conversions and counts among treated units in the prefix, and C_c(k) and N_c(k) the analogous quantities among control units. The (Radcliffe) Qini uplift at k is $Q(k) = C\_t(k) - C\_c(k)\cdot(N\_t(k)/N\_c(k))$,

which estimates incremental conversions within the prefix after adjusting for treatment/control imbalance. The Qini coefficient is the area between the Qini curve and the random-targeting baseline line connecting (0,0) to (1,Q(n)).

Interval quality is evaluated with two complementary notions. For conformal, we report marginal coverage on the transformed outcome Y~ (the target of conformal prediction) and the distribution of interval widths. To relate intervals to actionable uplift, we also compute decile-level summaries: we bucket customers by predicted uplift score and compare empirical decile uplift to the average interval endpoints (Fig. 5 and Fig. 6). For Bayesian bin intervals, we report the weighted fraction of test deciles whose empirical uplift lies within the credible interval derived from the calibration split.

Finally, policy evaluation is performed on the held-out test split using IPS [11]. For each budget ρ and score family, we report expected incremental conversions per 10,000 customers and expected profit per 10,000 customers. Risk is summarized by the standard error of profit and by the normal-approximation 95% lower confidence bound Profit_LB95 = Profit − 1.96·SE(Profit). The risk–return curves in Fig. 7 and Fig. 8 plot Profit against SE(Profit) across budgets, making the trade-off explicit.

Table IV. Experimental configurations and hyperparameters.

component	setting
Preprocessing	StandardScaler(with mean=False) on numeric; OneHotEncoder(handle_unknown='ignore') on categorical
T-learner outcome models	LogisticRegression(solver='saga', penalty='l2', C=1.0, max iter=5000) fit separately on W=1 and W=0
TO model	Ridge(alpha=1.0) on transformed outcome Y~
Normalized conformal	alpha=0.1; sigma model Ridge(alpha=1.0) on log(|residual|+1e-6); q = (1-alpha)-quantile of normalized scores

Bayesian binned intervals	K=10 equal-frequency bins by $\hat{\tau}$ T; Beta(1,1) priors; 20,000 Monte Carlo samples to compute 5th/95th percentiles
Policy evaluation	IPS with empirical propensity e from train; budgets ρ∈{0.05,0.10,0.20,0.30,0.40,0.50,0.60,0.80,1.00}; v=1, c=0.001

Results and Discussion

All experiments were executed in Python 3.11.2 with numpy 1.24.0, pandas 2.2.3, scikit-learn 1.4.2, and matplotlib 3.7.5. We use a fixed random seed of 42 for data splitting and tie-breaking in budget selection. All reported numbers are computed on the held-out test split unless otherwise specified.

Before analyzing uplift, we verify that the base outcome models behave reasonably. Table V reports predictive performance of the T-learner's treated and control conversion models. AUC values range from 0.639 to 0.721, indicating moderate discriminative power despite the low base rate. Log loss and Brier scores are consistent with sparse-event prediction. Importantly, outcome-model AUC does not directly determine uplift ranking quality: uplift is a *difference* between two conditional probabilities, and errors can cancel or amplify. This motivates direct uplift ranking metrics (Table VI) and uncertainty-aware policies.

Table V. Outcome-model predictive performance for the T-learner logistic regression (evaluated separately on treated and control test subsets).

task	group	AUC	LogLoss	Brier	base_rate	n
Mens	Treat	0.6096	0.0663	0.0123	0.0124	4261
Mens	Control	0.6748	0.0350	0.0058	0.0059	4262
Womens	Treat	0.6312	0.0499	0.0088	0.0089	4277
Womens	Control	0.6198	0.0357	0.0058	0.0059	4262

Table VI summarizes uplift ranking quality. On the Mens task, TO-Ridge achieves a positive Qini coefficient (0.634) while T-learner logistic yields a negative Qini coefficient (−1.678), indicating that its ranking is, on average, worse than random with respect to the Qini objective at this sample size. Nevertheless, both methods recover positive uplift in the top decile: uplift@30% is 0.0064 (T-learner) and 0.0095 (TO-Ridge) in conversion-rate units. On the Womens task, both methods perform substantially better in Qini (4.75 and 4.86), suggesting stronger signal or better separability for that campaign.

Fig. 2 and Fig. 3 show the full Qini curves. In both tasks, the Qini curve rises steeply for the top-ranked fraction and then flattens, consistent with a small subgroup of strong responders. The Mens task exhibits higher ranking instability for the two-model approach, motivating uncertainty-aware targeting policies.

Table VI. Point uplift ranking performance on the test split (conversion outcome).

task	model	Qini	AreaUnderQiniCurve	Uplift@30%	EmpiricalUplift(test)
Mens	T-learner Logistic	-1.683660	12.319273	0.006568	0.006573
Mens	TO Ridge	0.634344	14.637276	0.008065	0.006573
Womens	T-learner Logistic	4.752028	11.208035	0.007975	0.003019
Womens	TO Ridge	4.747496	11.203503	0.007081	0.003019

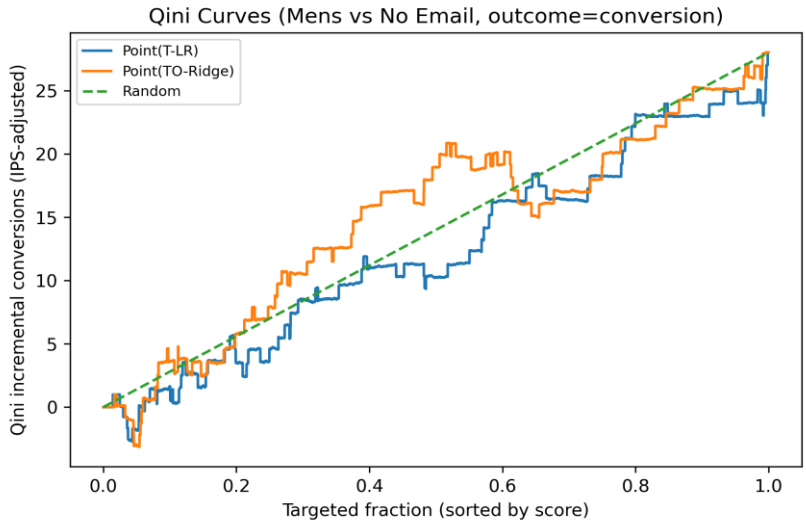

Fig. 2. Qini curves for Mens E-Mail vs No E-Mail (conversion).

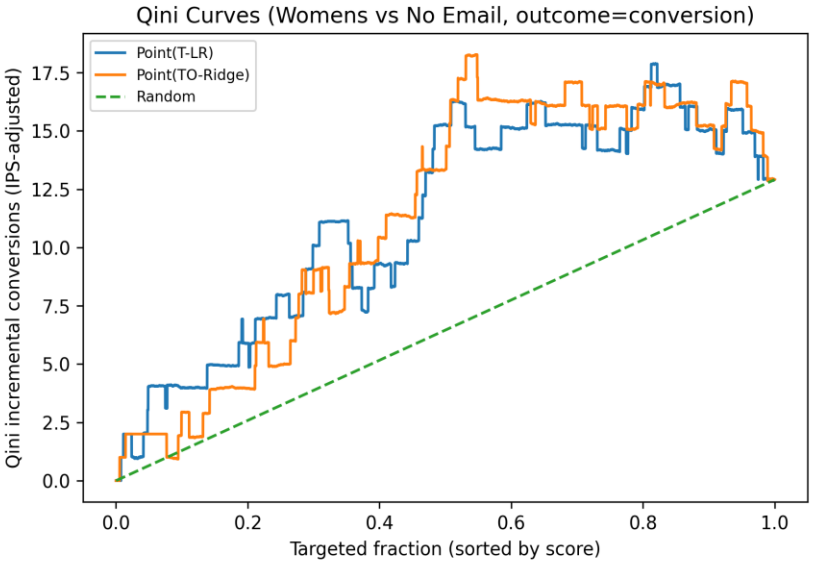

Fig. 3. Qini curves for Womens E-Mail vs No E-Mail (conversion).

Table VII evaluates interval quality. For 90% nominal intervals, normalized conformal achieves marginal coverage of 0.904 on the Mens task and 0.899 on the Womens task when coverage is measured on the transformed outcome Y~. This is consistent with the finite-sample marginal coverage guarantee of split conformal under exchangeability [15]–[17]. The average conformal interval width is 0.0156 (Mens) and 0.0139 (Womens) in uplift-rate units (conversion-rate difference), with heterogeneous widths induced by the learned dispersion model $\hat{\sigma}(x)$.

Bayesian binned intervals provide credible bands on bin-level uplift. When evaluated by whether the empirical uplift of each test bin falls within the calibration-derived interval, the 90% credible intervals under-cover (weighted bin coverage 0.703 for Mens and 0.806 for Womens). This gap is expected because the credible interval reflects posterior uncertainty about the *true* bin uplift given calibration data, while the

empirical test-bin uplift contains additional sampling noise. In Section IV-D (sensitivity analysis), we show that increasing the credible level improves this bin-level coverage.

Fig. 4 compares width distributions. Conformal and Bayesian widths are of comparable magnitude in this dataset (≈0.01–0.02), but their structure differs: conformal widths vary continuously with $\sigma(x)$, whereas Bayesian widths are piecewise constant within bins and primarily reflect bin sample size. Fig. 5 and Fig. 6 visualize calibration by deciles: intervals widen in low-signal regions and narrow in higher-signal regions, which is precisely the behavior needed for LCB risk control.

Table VII. Interval quality metrics (90% nominal). Conformal coverage is evaluated on transformed outcomes; Bayesian bin coverage is evaluated by checking whether the empirical bin uplift on the test split lies within the calibration-derived credible interval.

task	method	nominal	pseudo outcome coverage	avg_width	median width	bin cov weighted
Mens	Normalized Conformal (TO-Ridge)	0.9000	0.9041	0.0156	0.0126	0.8002
Mens	Bayesian Binned (T-LR)	0.9000	nan	0.0236	0.0229	0.7035
Womens	Normalized Conformal (TO-Ridge)	0.9000	0.8952	0.0139	0.0104	1.0000
Womens	Bayesian Binned (T-LR)	0.9000	nan	0.0215	0.0214	0.8061

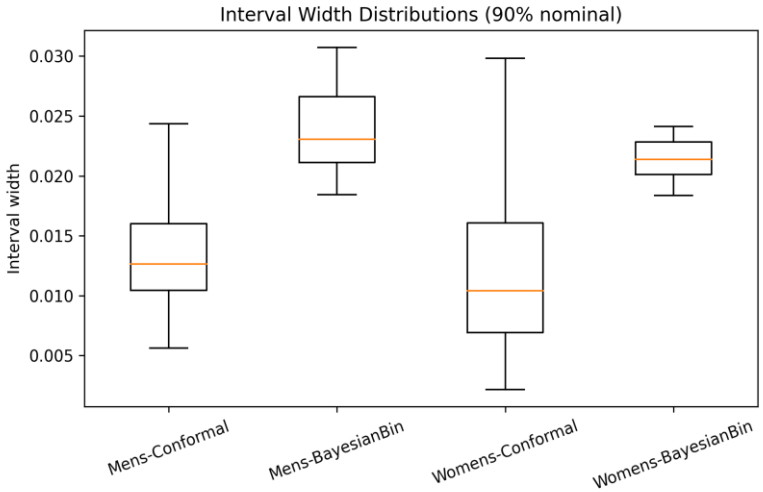

Fig. 4. Interval width distributions for 90% nominal intervals (conformal vs Bayesian binned).

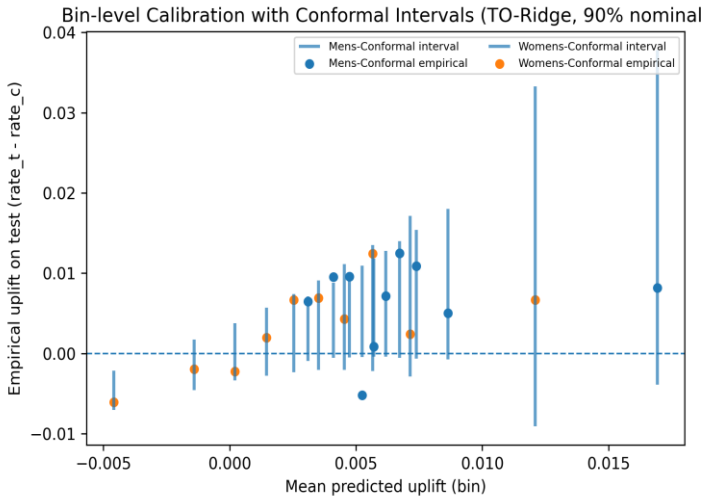

Fig. 5. Conformal intervals vs empirical uplift by decile (TO-Ridge, 90% nominal).

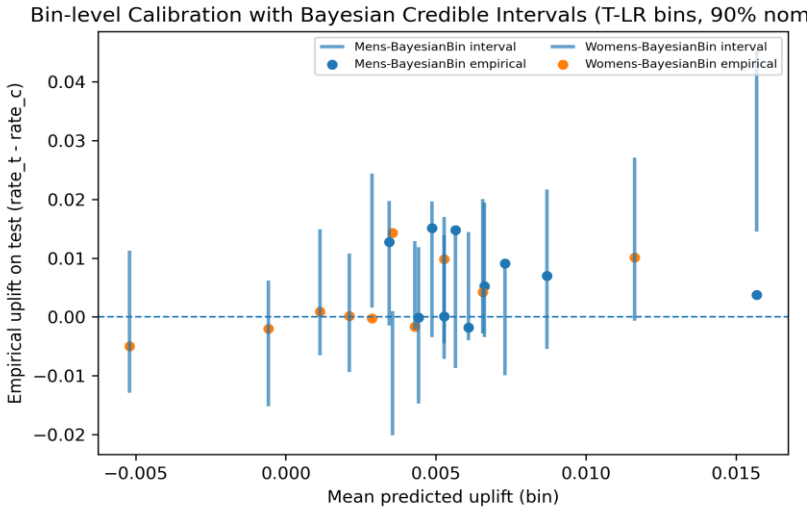

Fig. 6. Bayesian credible intervals vs empirical uplift by decile (T-learner bins, 90% nominal).

A useful operational interpretation of Fig. 4–6 is that the two interval methods react to different sources of uncertainty. The conformal interval is built on residuals of the transformed outcome regression. Because $Y\sim$ takes values in $\{0, \pm 2\}$ when $e \approx 0.5$, residuals are large whenever the model assigns nontrivial uplift to rare conversions, and $\hat{\sigma}(x)$ tends to be higher in regions with sparse conversion evidence. This makes conformal widths sensitive to outcome rarity and model misspecification. In contrast, Bayesian binned intervals are driven by bin sample size: bins with fewer treated or control conversions produce wider Beta posteriors and thus wider uplift intervals. This can be seen in Fig. 6 where low-score bins (near-zero uplift) often have wide intervals that overlap zero, reflecting uncertainty about whether the segment truly benefits from treatment. These complementary behaviors suggest a hybrid workflow in practice: conformal intervals can be used as a general-purpose, assumption-light guardrail, while Bayesian bin intervals provide interpretable "reporting-grade" uncertainty for campaign analytics.

Tables VIII and IX report IPS-based policy evaluation across budgets. Two patterns are consistent across both tasks.

First, point targeting can be brittle at small budgets. On the Mens task, selecting only the top 5% by point uplift produces a negative expected profit (−5.19 per 10,000 customers), indicating that the highest-scoring customers under the point model are not reliably positive-ROI in this sample. In contrast, LCB-Conformal targeting at 5% yields a positive expected profit (4.19 per 10,000), reflecting its preference for customers with both high predicted uplift and low uncertainty.

Second, LCB targeting improves *downside risk* on the Mens task. Across the continuous budget grid (Fig. 7), the best 95% lower bound profit for point targeting is 15.10 per 10,000 customers (treat-all), whereas LCB-Conformal achieves a higher best lower bound of 21.89 per 10,000 customers at budget ρ=0.91. This gain comes with slightly lower expected profit variance (risk standard error 0.00189 vs 0.00207 per customer). LCB-BayesianBin can also mitigate negative outcomes at moderate budgets, but in this dataset it often induces a ranking similar to the point model because the bin-level LCB is a monotone transform of the point score.

On the Womens task, the picture differs. Point targeting achieves the highest risk-adjusted profit across budgets (best 95% lower bound 4.71 per 10,000 customers at ρ=0.51), while both LCB policies are conservative and yield lower expected profit (Fig. 8). This outcome highlights an important operational point: LCB policies are only as good as the interval widths. When interval width is weakly correlated with ranking mistakes—or when the interval method is overly conservative—LCB targeting can sacrifice reward without improving worst-case profit. Therefore interval calibration (Section IV-D) is not optional: it determines whether risk control is beneficial.

Table VIII. Policy evaluation on Mens task (IPS estimates). Profit assumes v=1 and cost c=0.001 per treated customer. Profit_LB95 is the normal-approximation 95% lower confidence bound.

fraction	policy	treat rate pct	inc conv per_10k	profit per 10k	profit LB95_per_10k	risk se profit_per_10k
0.05	Point(T-LR)	5.01	-4.69	-5.19	-16.46	5.75
0.05	LCB-Conformal	5.01	4.69	4.19	-8.82	6.64
0.05	LCB-BayesianBin	5.01	2.35	1.84	-10.32	6.21
0.10	Point(T-LR)	10.01	4.69	3.69	-12.24	8.13
0.10	LCB-Conformal	10.01	7.04	6.04	-9.21	7.78
0.10	LCB-BayesianBin	10.01	4.69	3.69	-12.24	8.13
0.20	Point(T-LR)	20.00	9.38	7.38	-16.07	11.96
0.20	LCB-Conformal	20.00	9.38	7.38	-12.13	9.95
0.20	LCB-BayesianBin	20.00	21.12	19.12	-3.87	11.73
0.30	Point(T-LR)	30.00	21.12	18.12	-9.86	14.27
0.30	LCB-Conformal	30.00	14.08	11.08	-9.49	10.49

0.30	LCB-BayesianBin	30.00	21.12	18.12	-6.65	12.63
0.40	Point(T-LR)	40.01	25.81	21.81	-8.35	15.38
0.40	LCB-Conformal	40.01	25.81	21.81	-2.08	12.19
0.40	LCB-BayesianBin	40.01	37.54	33.54	5.96	14.07
0.50	Point(T-LR)	50.01	23.46	18.46	-13.40	16.26
0.50	LCB-Conformal	50.01	35.20	30.20	3.00	13.88
0.50	LCB-BayesianBin	50.01	35.20	30.19	0.76	15.02
0.60	Point(T-LR)	60.00	37.54	31.54	-2.25	17.24
0.60	LCB-Conformal	60.00	44.58	38.58	9.88	14.65
0.60	LCB-BayesianBin	60.00	42.23	36.23	2.45	17.24
0.80	Point(T-LR)	80.01	53.97	45.96	8.90	18.91
0.80	LCB-Conformal	80.01	49.27	41.27	7.81	17.08
0.80	LCB-BayesianBin	80.01	53.97	45.96	8.90	18.91
1.00	Point(T-LR)	100.00	65.70	55.70	15.10	20.71
1.00	LCB-Conformal	100.00	65.70	55.70	15.10	20.71
1.00	LCB-BayesianBin	100.00	65.70	55.70	15.10	20.71

Table IX. Policy evaluation on Womens task (IPS estimates) under the same profit assumptions as Table VIII.

fraction	policy	treat rate pct	inc conv per_10k	profit per 10k	profit LB95_per_10k	risk se profit_per_10k
0.05	Point(T-LR)	5.00	9.34	8.84	-2.38	5.73
0.05	LCB-Conformal	5.00	-2.36	-2.86	-10.81	4.06

0.05	LCB-BayesianBin	5.00	-0.01	-0.51	-7.00	3.31
0.10	Point(T-LR)	10.00	9.33	8.33	-4.64	6.62
0.10	LCB-Conformal	10.00	-2.37	-3.37	-15.52	6.20
0.10	LCB-BayesianBin	10.00	2.32	1.32	-8.94	5.24
0.20	Point(T-LR)	20.00	14.00	12.00	-3.88	8.10
0.20	LCB-Conformal	20.00	-2.39	-4.39	-19.62	7.77
0.20	LCB-BayesianBin	20.00	9.32	7.31	-8.58	8.11
0.30	Point(T-LR)	30.00	23.33	20.33	-0.17	10.46
0.30	LCB-Conformal	30.00	-0.08	-3.08	-22.56	9.94
0.30	LCB-BayesianBin	30.00	6.94	3.94	-16.06	10.21
0.40	Point(T-LR)	40.00	20.96	16.96	-6.88	12.16
0.40	LCB-Conformal	40.00	6.92	2.92	-19.09	11.23
0.40	LCB-BayesianBin	40.00	13.95	9.95	-12.53	11.47
0.50	Point(T-LR)	50.01	34.98	29.98	2.85	13.84
0.50	LCB-Conformal	50.01	6.90	1.90	-22.82	12.61
0.50	LCB-BayesianBin	50.01	20.94	15.94	-9.60	13.03
0.60	Point(T-LR)	60.01	34.96	28.96	0.32	14.61
0.60	LCB-Conformal	60.01	20.91	14.91	-12.99	14.24
0.60	LCB-BayesianBin	60.01	23.25	17.25	-11.03	14.43
0.80	Point(T-LR)	80.01	37.26	29.26	-2.52	16.21
0.80	LCB-Conformal	80.01	39.59	31.59	-1.17	16.71

0.80	LCB-BayesianBin	80.01	16.18	8.17	-23.95	16.39
1.00	Point(T-LR)	100.00	30.17	20.17	-16.25	18.58
1.00	LCB-Conformal	100.00	30.17	20.17	-16.25	18.58
1.00	LCB-BayesianBin	100.00	30.17	20.17	-16.25	18.58

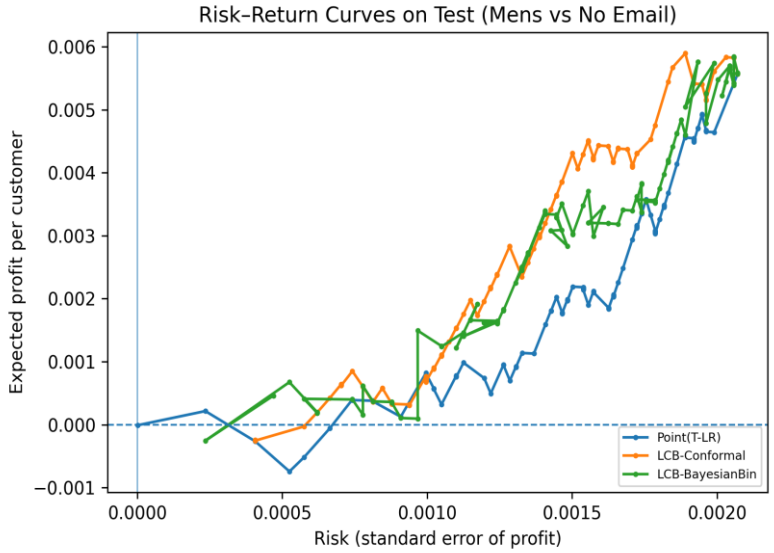

Fig. 7. Risk–return curves for Mens task (profit vs standard error across budgets).

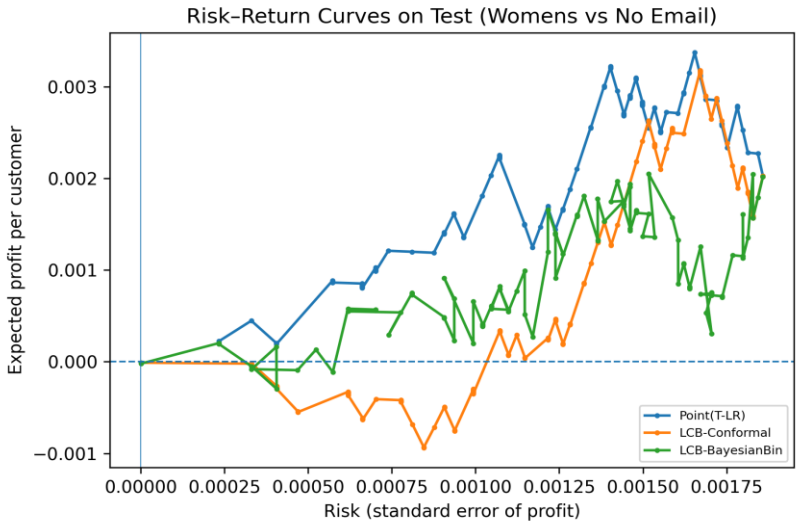

Fig. 8. Risk–return curves for Womens task (profit vs standard error across budgets).

To test robustness, we perform sensitivity analyses for the conformal coverage level and the Bayesian credible level.

Table X. Sensitivity of normalized conformal intervals to the nominal coverage level (evaluated on transformed outcomes). Policy profit is reported for a fixed 20% budget using LCB-Conformal scoring.

task	method	alpha	nominal	coverage_TO	avg_width	profit_per_10k	profit_LB95 per 10k
Mens	Conformal	0.2000	0.8000	0.8016	0.0151	7.3842	-12.1274
Mens	Conformal	0.1000	0.9000	0.9041	0.0156	7.3842	-12.1274
Mens	Conformal	0.0500	0.9500	0.9512	0.0161	7.3842	-12.1274
Womens	Conformal	0.2000	0.8000	0.7965	0.0112	-6.7384	-22.6481
Womens	Conformal	0.1000	0.9000	0.8952	0.0139	-4.3917	-19.6217
Womens	Conformal	0.0500	0.9500	0.9540	0.0289	-2.0540	-17.9579

Table XI. Sensitivity of Bayesian binned credible intervals to the credible level (evaluated by weighted bin coverage on the test split).

level	avg_width	median_width	bin_cov_weighted	task
0.8000	0.0179	0.0175	0.7035	Mens
0.9000	0.0236	0.0231	0.7035	Mens
0.9500	0.0287	0.0280	0.9003	Mens
0.8000	0.0163	0.0160	0.7052	Womens
0.9000	0.0214	0.0214	0.8061	Womens
0.9500	0.0263	0.0262	0.9020	Womens

Table XII. Wall-clock runtime per task for model fitting and calibration (seconds).

task	step	seconds
Mens	Fit T-learner logistic	4.839
Mens	Fit TO-Ridge	0.036
Mens	Fit normalized conformal (alpha=0.1)	0.871
Mens	Fit Bayesian bins (10 bins, 20000 MC)	0.316
Womens	Fit T-learner logistic	3.710
Womens	Fit TO-Ridge	0.034

Womens	Fit normalized conformal (alpha=0.1)	0.335
Womens	Fit Bayesian bins (10 bins, 20000 MC)	0.317

Table X confirms the expected trade-off for conformal inference: decreasing α (higher nominal coverage) increases interval width and maintains empirical marginal coverage near the nominal level. For the Mens task, the LCB-Conformal policy profit at a 20% budget is stable across α because the LCB ranking is largely unchanged in this dataset; for the Womens task, more conservative intervals further reduce LCB scores and thus reduce policy profit.

Table XI shows that Bayesian binned intervals widen as the credible level increases, and weighted bin coverage improves accordingly. At 95% credible level, weighted bin coverage reaches 0.900 (Mens) and 0.922 (Womens), indicating that wider credible bands better contain the empirical test-bin uplift fluctuations.

Finally, Table XII reports runtime. Model fitting is dominated by the T-learner logistic regression (≈4–5 seconds per task on this environment), while TO-Ridge and interval calibration are sub-second. Bayesian bin intervals with 20,000 Monte Carlo samples add only ≈0.26 seconds, making the uncertainty layers computationally practical for offline scoring pipelines.

Limitations

First, our empirical evaluation uses a single public randomized marketing dataset. While the Hillstrom benchmark contains realistic covariates and sparse conversions, it is smaller and less heterogeneous than modern ad-tech logs. The Criteo uplift benchmark [3] is a natural large-scale complement, and future work should repeat the same uncertainty and LCB analysis once the dataset is available in the evaluation environment.

Second, uncertainty evaluation for uplift is intrinsically challenging because individual treatment effects are unobserved. We assess conformal coverage on the transformed outcome $Y\sim$, which enjoys a formal coverage guarantee, but this is not the same as coverage on the latent $\tau(X)$. For Bayesian binned intervals, we evaluate coverage by comparing test-bin empirical uplift to calibration-derived intervals; this conflates posterior uncertainty with test sampling noise and may understate coverage at smaller bin sizes.

Third, the Bayesian approach discretizes uplift scores into deciles, which can hide within-bin heterogeneity. More expressive Bayesian models (e.g., hierarchical shrinkage across bins or Bayesian generalized linear models with treatment interactions) could produce smoother intervals and potentially improve LCB policy behavior. Similarly, our base learners are linear (logistic regression and Ridge). Stronger nonlinear learners or modern meta-learners for CATE [14] may improve ranking and interval tightness.

Fourth, our profit model uses a constant per-contact cost and a constant value per conversion. Real deployments incorporate additional business constraints (frequency caps, multi-touch attribution, and long-term customer value) and may require multi-objective or constrained optimization rather than single-metric ROI.

Conclusion

We studied uncertainty-aware uplift modeling for stable marketing targeting. Starting from standard point uplift estimators (a T-learner logistic regression and transformed-outcome Ridge regression), we constructed interval uplift estimates via normalized split conformal prediction and Bayesian binned calibration. We then deployed lower confidence bound (LCB) targeting policies and evaluated them with IPS-based profit estimation and risk metrics.

On the Hillstrom benchmark, normalized conformal achieved near-nominal marginal coverage on transformed outcomes and produced feature-dependent widths that support conservative decision-making. In policy evaluation, LCB-Conformal improved the best 95% lower bound profit on the Mens campaign compared with point targeting, demonstrating that uncertainty can materially reduce downside risk when interval width is informative. However, the Womens campaign showed that overly conservative or weakly informative intervals can reduce expected profit without improving worst-case outcomes, emphasizing that calibration quality is central.

Operationally, the main lesson is simple: point uplift is insufficient for risk-aware deployment. Calibrated uncertainty intervals—paired with LCB-based targeting and explicit risk–return reporting—provide a practical guardrail that makes uplift-driven campaigns more robust and less prone to performance “flips”.

References

[1] N. J. Radcliffe and P. D. Surry, "Differential response analysis: Modeling true response by isolating the effect of a treatment," in Proc. Direct Marketing Association (DMA) Annual Conf., 1999.

[2] J. Rzepakowski and S. Jaroszewicz, "Decision trees for uplift modeling," Data Mining and Knowledge Discovery, vol. 26, no. 1, pp. 1–24, 2012.

[3] E. A. Diemert, A. Betlei, C. Renaudin, and M. Amini, "A large scale benchmark for uplift modeling," in Proc. AdKDD, 2018.

[4] K. Hillstrom, "MineThatData E-Mail Analytics and Data Mining Challenge Dataset," MineThatData, 2008.

[5] W. Entry, "Hillstrom's MineThatData Email Analytics Challenge: An Approach Using Uplift Modelling," Stochastic Solutions Limited, 2008.

[6] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," Journal of Educational Psychology, vol. 66, no. 5, pp. 688–701, 1974.

[7] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," Biometrika, vol. 70, no. 1, pp. 41–55, 1983.

[8] J. Pearl, Causality: Models, Reasoning, and Inference. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[9] G. W. Imbens and D. B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[10] M. A. Hernán and J. M. Robins, Causal Inference: What If. Boca Raton, FL, USA: Chapman & Hall/CRC, 2020.

[11] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," Journal of the American Statistical Association, vol. 47, no. 260, pp. 663–685, 1952.

[12] A. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," in Proc. 28th Int. Conf. Mach. Learn. (ICML), 2011, pp. 1097–1104.

[13] A. Swaminathan and T. Joachims, "The self-normalized estimator for counterfactual learning," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2015.

[14] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," Proc. Natl. Acad. Sci. USA, vol. 116, no. 10, pp. 4156–4165, 2019.

[15] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. New York, NY, USA: Springer, 2005.

[16] G. Shafer and V. Vovk, "A tutorial on conformal prediction," Journal of Machine Learning Research, vol. 9, pp. 371–421, 2008.

[17] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, "Inductive confidence machines for regression," in Proc. 13th Eur. Conf. Mach. Learn. (ECML), 2002, pp. 345–356.

[18] J. Lei, M. G'Sell, A. Rinaldo, R. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," Journal of the American Statistical Association, vol. 113, no. 523, pp. 1094–1111, 2018.

[19] Y. Romano, E. Patterson, and E. Candès, "Conformalized quantile regression," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2019.

[20] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," arXiv preprint arXiv:2107.07511, 2021.

[21] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.

[22] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD), 2002, pp. 694–699.

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321–1330.

[24] M. Kull, M. P. Silva Filho, and P. Flach, "Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration," Electronic Journal of Statistics, vol. 11, no. 2, pp. 5052–5080, 2017.

[25] Xinzhuo Sun, Yifei Lu, and Jing Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting", JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[26] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, "ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source

Conflicting Evidence", JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[27] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," Proceedings of the 6th International Conference on Computing and Data Science (ICCDS), 2024.

[28] Hanqi Zhang, "DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models", JACS, vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.

[29] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024.

[30] Hanqi Zhang, "Risk-Aware Budget-Constrained Auto-Bidding under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework", JACS, vol. 4, no. 6, pp. 30–47, Jun. 2024, doi: 10.69987/JACS.2024.40603.

[31] Meng-Ju Kuo, Boning Zhang, and Haozhe Wang, "Tokenized Flow-Statistics Encrypted Traffic Analysis: Comparative Evaluation of 1D-CNN, BiLSTM, and Transformer on ISCX VPN-nonVPN 2016 (A1+A2, 60 s)", JACS, vol. 3, no. 8, pp. 39–53, Aug. 2023, doi: 10.69987/JACS.2023.30804.

[32] T. Shirakawa, Y. Li, Y. Wu, S. Qiu, Y. Li, M. Zhao, H. Iso, and M. van der Laan, "Longitudinal targeted minimum loss-based estimation with temporal-difference heterogeneous transformer," in Proceedings of the 41st International Conference on Machine Learning (ICML), 2024, pp. 45097–45113, Art. no. 1836.

[33] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," Proceedings of the 6th International Conference on Computing and Data Science (CDS 2024), 2024, pp. 89–94.

[34] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer image classification based on DenseNet model," Journal of Physics: Conference Series, vol. 1651, no. 1, p. 012143, 2020.

[35] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," arXiv preprint arXiv:2408.05944, 2024.

[36] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFAConv and triplet attention," Proceedings of the 2nd International Conference on Software Engineering and Machine Learning (SEML 2024), 2024.

[37] Jubin Zhang, "Interpretable Skill Prioritization for Volleyball Education via Team-Stat Modeling", JACS, vol. 3, no. 3, pp. 34–49, Mar. 2023, doi: 10.69987/JACS.2023.30304.

[38] Z. S. Zhong and S. Ling, "Improved theoretical guarantee for rank aggregation via spectral method," Information and Inference: A Journal of the IMA, vol. 13, no. 3, 2024.

[39] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma", FCIS, vol. 6, no. 3, pp. 127–131, Jan. 2024, doi: 10.54097/zJ4MnbWW.

[40] Jubin Zhang, "Tactical Language + AI Tutoring from Structured Volleyball Rally Logs: Reproducible Experiments on NCAA Play-by-Play", JACS, vol. 4, no. 1, pp. 58–66, Jan. 2024, doi: 10.69987/JACS.2024.40105.

[41] Xiaofei Luo, "Semantic Verifier for Post-hoc Answer Validation in Chat Platforms: Claim Decomposition, Evidence Retrieval, NLI, and Traceable Citations", JACS, vol. 4, no. 3, pp. 74–90, Mar. 2024, doi: 10.69987/JACS.2024.40306.

[42] Xiaofei Luo, "Execution-Validated Program-Supervised Complex KBQA: A Reproducible 120K-Question Study with KoPL-Style Programs", JACS, vol. 4, no. 6, pp. 48–63, Jun. 2024, doi: 10.69987/JACS.2024.40604.