Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 17-38
DOI: https://doi.org/10.51903/jtie.v5i1.468

Privacy-Robust Incrementality Estimation in Cookieless Settings via Uplift Modeling: Reproducible Evidence from the Hillstrom E-Mail Experiment

Jingwen Bai*[1], Haozhe Wang[2], Qiyou Wu[3], Boning Zhang[4]
Email*: jwbai1118@outlook.com
[1]*Data Science, Columbia University, NY, USA*
[2]*Operations Research and Information Engineering, Cornell, NY, USA*
[3]*Artificial Intelligence, Northeastern University, MA, USA*
[4]*Computer Science, Georgetown University, DC, USA*
*Corresponding Author

Abstract
Measuring advertising incrementality without user-level identifiers is increasingly constrained by platform policies and privacy regulations. In cookieless environments, practitioners often observe only aggregated signals (e.g., cohort-level conversion counts) yet must estimate causal lift and quantify uncertainty. This paper studies cookieless incrementality evaluation through uplift (ITE/CATE) modeling under explicit privacy constraints. Using the public MineThatData (Hillstrom) E-Mail Analytics Challenge randomized experiment (64,000 customers; three arms), we cast the task as a binary treatment problem—sending any e-mail versus sending none—and compare six ITE estimators (S-, T-, X-, R-, and doubly robust learners, plus transformed-outcome regression) with cohort-only estimators that emulate privacy-preserving reporting. Cohort estimation uses only aggregated counts with a Bayesian beta–binomial shrinkage model, and we evaluate robustness under k-anonymity suppression and Laplace-noised differentially private aggregates. On held-out test data, the best ID-level model (T-learner with logistic regression) achieves a Qini coefficient of 6.675 and improves the estimated policy conversion rate when targeting the top 20% by predicted uplift. Cohort-only estimation retains a weaker, higher-variance signal; its point estimate is sensitive to privacy settings but provides uncertainty intervals with 0.892 empirical coverage for a nominal 95% cohort-level interval. Overall, the results show that (i) causal lift remains estimable under randomized experiments without identifiers, (ii) robust meta-learners offer strong performance and fast scoring, and (iii) aggregation and privacy noise introduce a quantifiable accuracy–privacy trade-off.

Keywords: *Cookieless Measurement, Incrementality, Uplift Modeling, Heterogeneous Treatment Effects, Differential Privacy.*

I. INTRODUCTION

Incrementality—the causal effect of an advertising intervention relative to a counterfactual without the intervention—is the core quantity needed for budget allocation, bidding, and experimentation-driven growth. In practice, many marketing organizations still rely on observational attribution (e.g., last-touch or multi-touch heuristics) because it is cheap to compute from event logs. Yet attribution is not incrementality: it answers 'who was observed before a conversion,' not 'what would have happened without the intervention.' When campaigns affect user behavior and measurement itself is selective, attribution can be systematically biased. Randomized controlled trials (RCTs) remain the clearest path to credible incrementality (Kohavi et al., 2009; Rubin, 1974). While the motivating application is advertising measurement, the core contribution of this manuscript is methodological—privacy-aware data processing, AI/ML-based

Received on December, 2025; Revised on January, 2026; Accepted on February, 2026; Published on April 2026

causal estimation, and uncertainty-aware interpretation for aggregated measurement pipelines—rather than marketing strategy or business analytics per se.

Modern measurement faces a structural shift: cross-site identifiers and third-party cookies are restricted, while platform APIs increasingly return aggregated or delayed reports designed to protect user privacy (Apple, 2021; Google, 2025). In these settings, the measurement pipeline becomes 'cookieless': analysts may not be able to match impressions to conversions at the person level. Instead, they must work with cohorts, summary statistics, and weak signals. A practical consequence is that common monitoring questions—'did campaign A cause incremental conversions?' and 'which audiences should we target?'—must be answered with less granular data.

Two technical challenges follow: first, lift estimation must remain statistically stable even when the analyst only observes aggregated outcomes and covariates. Aggregation reduces variance by pooling but also removes information needed for personalization. Second, uncertainty must be quantified; without reliable uncertainty bands (Sun et al., 2024), lift estimates become brittle inputs to decision systems. This is especially important when measurement APIs apply privacy mechanisms that introduce additional randomness. Uplift modeling—originally developed for targeted marketing—directly addresses these challenges by estimating heterogeneous treatment effects (HTEs) or individual treatment effects (ITEs) (Lo, 2002; Radcliffe, 2007). If the experiment assignment mechanism is known, uplift models can rank cohorts or individuals by expected incremental impact. The question is whether this approach remains reliable when identifiers are unavailable, and privacy constraints (e.g., k-anonymity, differential privacy) affect the data available for modeling.

Cookieless measurement is not only a data engineering constraint but also a statistical one. When only aggregated counts are observable, the analyst's estimand must be carefully defined, and the evaluation protocol must respect the reporting mechanism. For example, both Apple's SKAdNetwork and the Privacy Sandbox Attribution Reporting API expose conversion signals through aggregation and/or noise, limiting event-level linkage (Apple, n.d.; Google, 2025). This motivates two complementary goals: (i) design estimators that remain meaningful under aggregation and noise, and (ii) build diagnostics that quantify the accuracy–privacy trade-off so that practitioners can choose privacy parameters intentionally rather than implicitly.

This paper presents a reproducible empirical study of cookieless incrementality evaluation using a publicly available randomized marketing dataset. We use the MineThatData (Hillstrom) E-Mail Analytics Challenge dataset, which contains 64,000 customers randomly assigned to receive either a men's e-mail campaign, a women's e-mail campaign, or no e-mail (Hillstrom, 2008;

sklift, 2021). We collapse the two treatment arms into a single binary treatment ('any e-mail') and evaluate multiple ITE estimators. We then simulate cookieless measurement by restricting the analyst to cohort-level aggregates and applying privacy constraints: (i) k-anonymity thresholds that suppress small cohorts and (ii) differentially private Laplace noise added to cohort counts (Dwork et al., 2006; Sweeney, 2002). We report performance using uplift-specific metrics (Qini and AUUC) and off-policy policy value estimation using inverse propensity weighting (Radcliffe & Surry, 2011).

The empirical findings show a clear separation between ID-level ITE estimation and cohort-only estimation. The best ID-level approach in our experiments, a T-learner with logistic regression, achieves a Qini coefficient of 6.675 and the highest policy conversion rate among the methods compared. Doubly robust learning yields comparable Qini while scoring orders of magnitude faster, reflecting its appeal for production systems that require low-latency ranking. Cookieless cohort estimation is substantially noisier; nevertheless, Bayesian shrinkage and explicit uncertainty reporting provide a principled way to monitor degradation as privacy constraints intensify. These results support a pragmatic recommendation: when individual identifiers are absent, invest in experiment design, stable cohort definitions, and robust estimators with transparent uncertainty rather than relying on deterministic point estimates.

In summary, the paper makes three concrete contributions. (1) We provide a full experimental comparison of widely used uplift/CATE learners on a public randomized marketing dataset, including ranking metrics, policy value, and computational cost. (2) We operationalize a cookieless evaluation protocol that restricts the analyst to cohort-level aggregates and explicitly models privacy mechanisms (k-anonymity suppression and Laplace-noised differential privacy). (3) We quantify uncertainty in both ID-level and cookieless settings using bootstrap and Bayesian intervals, and we validate interval calibration empirically on a held-out split.

The remainder of the manuscript reviews related work on uplift modeling, causal inference, and privacy-preserving measurement; describes the dataset, models, and evaluation metrics; reports experimental results with detailed tables and figures; and concludes with practical recommendations for cookieless incrementality workflows.

II. LITERATURE REVIEW

Causal inference for incrementality is commonly framed through the potential outcomes model, where each unit has a potential outcome under treatment and control, and the causal effect is defined as their difference (Rubin, 1974). In online and marketing contexts, controlled experiments are widely viewed as the most credible approach to identifying causal effects, and practical guidance emphasizes randomization, guardrail metrics, and careful power analysis

(Kohavi et al., 2009). When randomization is infeasible, observational causal inference methods rely on assumptions such as ignorability and are typically more fragile (Hernán & Robins, 2020; Imbens & Rubin, 2015; Pearl, 2009).

From a decision-making perspective, incrementality problems can be formulated at multiple levels of granularity. The average treatment effect (ATE) summarizes the mean impact of treatment across the population, while the conditional average treatment effect (CATE) and individual treatment effect (ITE) describe how the impact varies with covariates. In marketing, this heterogeneity matters because budgets are constrained: an optimal policy may treat only units for which the expected lift exceeds the cost. The conceptual foundation is the same: for each unit $i$with covariates $\mathbf{X}_i$, treatment $T_i \in \{0,1\}$, and outcome Y_i, the causal effect is defined as $Y_i(1) - Y_i(0)$, which is not directly observable for any single unit. Identification, therefore, relies on assumptions such as randomized assignment (in our experiments) or unconfoundedness together with overlap (in observational settings) (Imbens & Rubin, 2015; Hernán & Robins, 2020). Practical complications include interference (one user's treatment affecting another's outcome) and noncompliance, which can be particularly relevant in advertising, where auctions and frequency capping can couple outcomes across users (Pearl, 2009).

Uplift modeling (Shirakawa et al., 2024) extends the causal framing from average treatment effects to heterogeneous treatment effects (HTEs), enabling targeting policies that treat only those units expected to respond positively. Early work in marketing introduced 'true lift' models and highlighted the need for control groups to estimate incremental response (Lo, 2002; Radcliffe, 2007). Subsequent research developed specialized learners such as uplift trees and ensemble methods, and provided surveys of algorithmic families and evaluation practices (Devriendt et al., 2018; Gutierrez & Gérardy, 2017; Rzepakowski & Jaroszewicz, 2012).

A line of work focuses on tree-based and ensemble learners that directly optimize splitting criteria for treatment-effect heterogeneity. Uplift trees extend decision tree induction by selecting splits that maximize estimated uplift rather than classification purity, and have been adapted to multiple treatments and continuous outcomes (Rzepakowski & Jaroszewicz, 2012). In the broader causal machine learning literature, causal forests and generalized random forests provide nonparametric estimators of CATE that are asymptotically normal under regularity conditions (Athey et al., 2019; H. Jamaludin et al., 2024; Wager & Athey, 2018). These methods are attractive when interactions are complex, but they typically require more data and careful tuning; in privacy-constrained settings, the variance introduced by noise and aggregation can interact with high-capacity learners in nontrivial ways.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 17-38
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

A major modern thread connects uplift modeling to the machine learning literature on conditional average treatment effects (CATEs). Meta-learners formalize how to leverage supervised learning algorithms for treatment effect estimation: S-learners model outcomes as a function of covariates and treatment; T-learners fit separate models for each treatment arm; X-learners improve performance under treatment imbalance by imputing unit-level effects and re-learning them; and R-learners reduce bias by residualizing outcomes and treatment (Künzel et al., 2019; Nie & Wager, 2021). Doubly robust (DR) estimators combine outcome models and propensity models and retain consistency if either nuisance model is correctly specified, which is valuable in practical settings with model mis-specification risk (Bang & Robins, 2005; Chernozhukov et al., 2018).

Doubly robust and debiased machine learning frameworks explicitly decompose the estimation task into nuisance functions and a target parameter of interest. In the context of conditional average treatment effect (CATE) estimation, the nuisance components comprise the outcome regression functions $\mu_t(\mathbf{X}) = \mathbb{E}[Y \mid T = t, \mathbf{X}]$and the propensity score $e(\mathbf{X}) = \mathbb{P}(T = 1 \mid \mathbf{X})$. To reduce overfitting-induced bias, cross-fitting and sample-splitting procedures are employed, ensuring that nuisance function estimates are evaluated on data not used during their training phase (Chernozhukov et al., 2018). Although the Hillstrom dataset is randomized (so e(X) is constant in expectation), the doubly robust perspective remains useful for cookieless settings because measurement systems can induce missingness, selection, or differential reporting that effectively behaves like a non-random observation process.

Evaluating uplift models (Zhang, 2025) differs from standard classification because the counterfactual outcome for each individual is unobserved. Ranking-based metrics summarize how well a model concentrates incremental gain among the top-scored units. The Qini curve and Qini coefficient, inspired by the Gini coefficient, are widely used and can be computed from randomized data by scaling control outcomes (Zhang, 2024) to the treated sample size (Radcliffe & Surry, 2011). Practical uplift evaluation also emphasizes policy value—the expected outcome under a targeting policy—often estimated using inverse propensity weighting or doubly robust off-policy estimators when the policy differs from the randomized assignment (Künzel et al., 2019).

Policy evaluation (Zhang, 2023) is especially important for incrementality because real decisions are policy-based: advertisers decide who to target, how often, and at what cost. When an evaluation policy differs from the logging policy (e.g., randomized assignment), off-policy estimators such as Inverse Propensity Scoring (IPS) and doubly robust estimators provide unbiased or low-bias estimates of policy value under standard assumptions (Künzel et al., 2019;

Hernán & Robins, 2020). In this paper, we use IPS with known randomization probability to estimate the expected conversion rate under targeting policies defined by predicted uplift rankings.

The cookieless shift introduces an additional dimension: privacy constraints affect what data can be observed and released. Privacy-preserving measurement systems increasingly provide aggregated and/or noisy reports; examples include Apple's SKAdNetwork for mobile app attribution and the Privacy Sandbox Attribution Reporting API for web measurement (Apple, n.d.; Google, 2025). From a data analysis perspective, this resembles an aggregation-and-noise mechanism applied to conversions, exposures, or other events. Classical privacy models include k-anonymity, which requires that each released record be indistinguishable from at least k individuals, and differential privacy (DP), which bounds the effect that any individual can have on the released output distribution (Dwork & Roth, 2014; Sweeney, 2002). A core implication is an accuracy–privacy trade-off: more privacy implies less precise measurement.

Differential privacy formalizes privacy as a property of a randomized algorithm: for neighboring datasets differing by a single individual, the output distributions are close up to a multiplicative factor $\exp(\varepsilon)$ (and optionally an additive δ). The Laplace mechanism achieves ε-differential privacy for numeric queries with bounded sensitivity by adding Laplace noise whose scale is proportional to the sensitivity divided by ε (Dwork et al., 2006; Dwork & Roth, 2014). In aggregate conversion reporting, sensitivity is often one (a single user can contribute at most one conversion), making Laplace-noised counts a natural abstraction. However, composition over many reports and post-processing steps can reduce effective accuracy, motivating empirical studies of robustness like the one conducted here.

Uncertainty quantification is therefore integral to cookieless incrementality. For scalar effects such as average treatment effects, frequentist confidence intervals based on asymptotic normality or bootstrap resampling are standard. For cohort-level conversion rates, Bayesian beta–binomial models provide analytically tractable posterior uncertainty, while bootstrapping can quantify the variability of ranking metrics such as Qini under repeated sampling (Efron & Tibshirani, 1994; Hernán & Robins, 2020). For cohort-level conversion rates, Bayesian beta–binomial models provide analytically tractable posterior uncertainty; for treatment-effect differences, posterior sampling can propagate uncertainty into uplift intervals. Bootstrapping can quantify the variability of ranking metrics such as Qini across repeated samples, which is important because ranking metrics are nonlinear functionals of the data and can be sensitive to small changes. Recent work has also emphasized the role of uncertainty in uplift in practical targeting, particularly under limited data and noise (Radcliffe & Surry, 2011). In privacy-preserving measurement, reporting

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 17-38
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

noise introduces an additional source of uncertainty that must be distinguished from sampling variability.

III. RESEARCH METHOD

Figure 1 summarizes the end-to-end evaluation pipeline used in this study, connecting the randomized experiment to both ID-level uplift modeling and cookieless, privacy-preserving cohort evaluation.

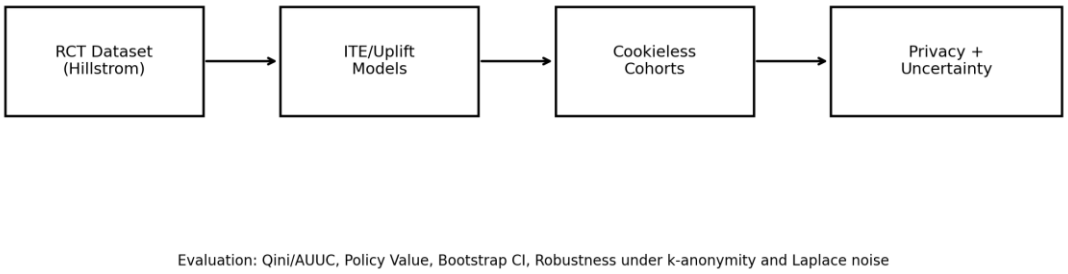


Figure 1. End-to-End Pipeline for ID-Level and Cookieless Uplift Evaluation

A. Dataset and task formulation

We use the MineThatData (Hillstrom) E-Mail Analytics Challenge dataset, which contains 64,000 customers who purchased within the last 12 months and were randomly assigned to one of three groups: Men's E-Mail, Women's E-Mail, or No E-Mail (Hillstrom, 2008; sklift, n.d.). We define the binary treatment indicator T as 1 for Men's E-Mail or Women's E-Mail and 0 for No E-Mail. The primary outcome Y is conversion (binary). We report visit and spend statistics for completeness, but use conversion as the uplift target.

B. Binary treatment definition

The original experiment contains two active treatments (Mens E-Mail and Womens E-Mail). In many cookieless advertising measurement systems, reporting is available only for a coarse treatment notion (e.g., 'campaign on' vs 'campaign off'), so we collapse the two e-mail arms into a single treatment. This yields a treatment-assignment probability of approximately 2/3 in the full dataset (Table 2). We retain the original segment label for descriptive analysis (Figure 2) to verify that both e-mail variants outperform the no-e-mail control in average conversion rate. This simplification necessarily discards potential treatment-effect heterogeneity between men's and women's creative variants; we therefore interpret the results as the incremental effect of 'any e-mail' relative to no e-mail, and treat the loss of arm-specific heterogeneity as a methodological limitation.

C. Train/test protocol and software

We perform a single 70/30 train–test split stratified by the binary treatment indicator (random seed 42). All reported ranking metrics and policy values are computed on the held-out test split and therefore reflect out-of-sample performance. We implemented the experiments in Python (Python 3.11) using pandas for data handling and scikit-learn (Pedregosa et al., 2011) for model training (logistic regression and ridge regression). The use of a fixed random seed, explicit hyperparameters (Table 6), and saved artifacts (tables and figures) ensures that the reported results are fully reproducible.

D. Experimental design and estimand

The original dataset is a three-arm randomized experiment. We focus on the binary estimand that compares sending any promotional e-mail with not sending any e-mail. This choice is motivated by cookieless measurement, in which treatments are often collapsed into an 'exposed' indicator when precise creative/placement attribution is unavailable. Collapsing two treatment arms increases the treated sample size and yields a stable estimate of the incremental effect of e-mail contact, while still preserving meaningful heterogeneity because customer covariates (e.g., purchase history and browsing channel) vary widely. In Table 2, each arm contains tens of thousands of customers, enabling both average- and heterogeneous-effect estimation.

E. Implementation details and reproducibility

We split the dataset into 70% training and 30% testing using stratification by the binary treatment indicator and a fixed random seed (seed=42). All reported results are computed on the held-out test split. Model training and evaluation were implemented in Python using pandas for data processing and scikit-learn (Pedregosa et al., 2011) for logistic and ridge regression. Because the data arise from randomized assignment, the treatment propensity is known; we use the empirical treatment rate in the training split as the propensity in policy evaluation and in doubly robust pseudo-outcomes.

Table 1. Hillstrom Dataset Schema Used in This Study

column	dtype	role
recency	int64	feature
history_segment	object	feature
history	float64	feature
mens	int64	feature
womens	int64	feature
zip_code	object	feature
newbie	int64	feature
channel	object	feature
segment	object	treatment
visit	int64	outcome
conversion	int64	outcome
spend	float64	outcome

Table 2. Descriptive Statistics by Original Experimental Segment (Full Dataset)

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 17-38
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

segment	n	visit_rate	conversion_rate	avg_spend
Mens E-Mail	21307	0.182757	0.012531	1.422617
No E-Mail	21306	0.106167	0.005726	0.652789
Womens E-Mail	21387	0.1514	0.008837	1.077202

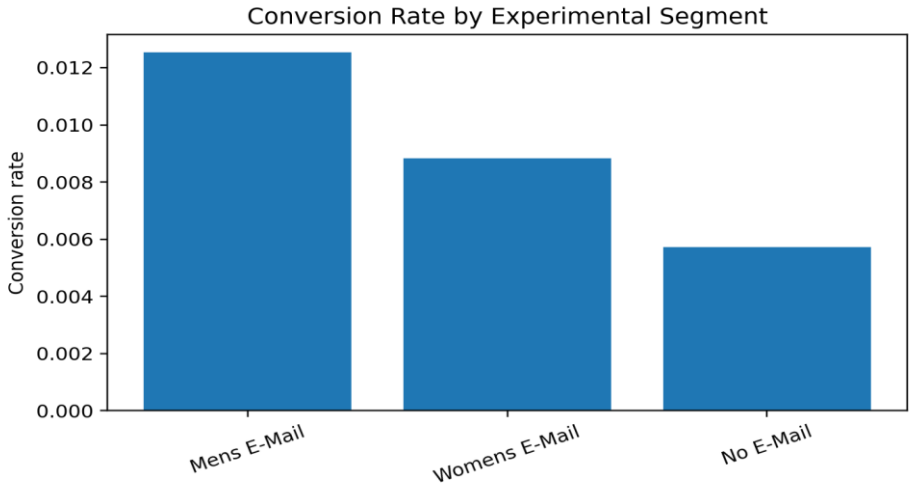


Figure 2. Conversion Rate by Original Experimental Segment

Table 3. Aggregated Outcomes for Binary Treatment (Any E-Mail vs No E-Mail)

group	n	conversion_rate	visit_rate	avg_spend
Control (No E-Mail)	21306	0.005726	0.106167	0.652789
Treatment (Any E-Mail)	42694	0.010681	0.167049	1.249585

Table 4. Average Treatment Effect (ATE) on Conversion with 95% Confidence Interval

metric	estimate	ci_95_low	ci_95_high	treat_rate	control_rate	n_treat	n_control
ATE (conversion rate difference)	0.004955	0.003548	0.006361	0.010681	0.005726	42694	21306

F. Preprocessing and feature representation

We one-hot encode categorical variables (zip_code, channel, and history_segment) and keep numeric/binary variables (recency, history, mens, womens, newbie). All features are standardized using the training split statistics. This results in a compact tabular representation suitable for linear ITE estimators.

G. Cookieless simulation

To emulate an environment without stable person-level identifiers, we restrict the analyst to cohort-level aggregates. Cohorts are defined by coarsening continuous variables into quantile bins and combining them with discrete attributes (zip_code, channel, history_segment, newbie). Importantly, the quantile cutpoints are computed on the training split and reused on the test split to create a stable cohort taxonomy. This mirrors real privacy-preserving measurement systems

that report only aggregated statistics keyed by a finite set of dimensions, where continuous attributes must be bucketed.

Table 5. Feature Encoding Summary (After One-Hot Encoding)

feature_group	count
numeric	5
categorical	3
one_hot_total	13
total_features	18

We use quantile binning to keep cohorts reasonably balanced, which reduces extreme variance in rate estimates and increases the likelihood that cohorts meet minimum cell-size constraints. Cohort granularity is a tunable design choice: coarser cohorting increases statistical stability but reduces heterogeneity, while finer cohorting increases resolution but may be more frequently suppressed (k-anonymity) or dominated by noise (DP). For each cohort and treatment arm, we compute only the counts (n, y), where n is the number of users and y is the number of conversions. Individual-level models are not allowed access to row-level outcomes in this cookieless variant.

H. Privacy constraints

We evaluate two privacy mechanisms: (1) k-anonymity, where we suppress cohorts whose total sample size falls below k_min; this mimics reporting systems that enforce minimum aggregation thresholds to reduce re-identification risk (Sweeney, 2002). (2) Differential privacy: we add Laplace noise to cohort exposure and conversion counts and then compute conversion rates from the noisy counts. We assume each individual contributes at most 1 to each released count, so the L1 sensitivity of each count query is 1, and the Laplace scale is $1/\varepsilon$. This emulates privacy-preserving aggregate reporting, in which outputs are perturbed by noise calibrated to a privacy budget ε (Dwork et al., 2006; Dwork & Roth, 2014). For simplicity, ε is treated as a per-count budget, and we intentionally abstract away composition across multiple released counts, so ε should be interpreted as controlling per-report noise magnitude in our simulation. In each privacy setting, we fit cohort-only estimators on the privatized training aggregates and evaluate them on the privatized test aggregates. Table 6 summarizes the privacy mechanisms and the stress-test parameters.

I. ITE models

We compare six ID-level ITE estimators and two cohort-only estimators. The ID-level estimators are: (i) S-learner with logistic regression and explicit interaction terms; (ii) T-learner with separate logistic regressions for treated and control; (iii) X-learner using logistic regression outcome models and ridge regression effect models; (iv) DR-learner using an augmented inverse propensity weighted (AIPW) pseudo-outcome and ridge regression; (v) R-learner using residualized outcomes/treatments and ridge regression; and (vi) transformed-outcome regression,

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 17-38
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

which is unbiased under randomized treatment assignment. The cohort-only estimators are: (vii) cohort raw difference-in-means; and (viii) cohort Bayesian shrinkage using independent beta–binomial posteriors per arm.

J. Model details (ID-level)

The S-learner fits a single outcome model μ(X,T) and then computes the uplift as μ(X,1)−μ(X,0). We implement μ using L2-regularized logistic regression and explicitly include interaction terms T×X, so that the linear model can capture heterogeneous treatment effects rather than only a global shift. The T-learner fits two separate logistic regressions, μ_1(X)=P(Y=1|T=1,X) and μ_0(X)=P(Y=1|T=0,X), and predicts uplift as μ_1(X)−μ_0(X). In randomized marketing data, T-learners often perform well as baselines because they can capture different covariate–outcome relationships across arms without imposing additivity.

The X-learner improves upon S- and T-learners under treatment imbalance by borrowing information across treatment arms through imputation (Künzel et al., 2019). We first estimate the outcome regressions $\mu_0(\mathrm{X})$and $\mu_1(\mathrm{X})$, and subsequently construct imputed unit-level treatment effects. For treated units, the imputed effect is defined as $D_1 = Y - \mu_0(\mathrm{X})$; for control units, $D_0 = \mu_1(\mathrm{X}) - Y$. Regression models are then fitted to approximate $\tau_1(\mathrm{X}) \approx \mathbb{E}[D_1 \mid \mathrm{X}, T = 1]$and $\tau_0(\mathrm{X}) \approx \mathbb{E}[D_0 \mid \mathrm{X}, T = 0]$, implemented using ridge regression. The final CATE estimator is obtained as $\tau(\mathrm{X}) = (1 - e)\tau_0(\mathrm{X}) + e\,\tau_1(\mathrm{X})$, where $e$denotes the (known) treatment propensity. In our setting, the binary treatment aggregates two treatment arms, yielding a propensity of approximately $2/3$. Although the assignment is randomized, this moderate imbalance renders the X-learner particularly appropriate.

K. Model details (robust learners)

The DR-learner constructs a doubly robust pseudo-outcome based on the augmented inverse propensity weighting (AIPW) identity (Bang & Robins, 2005). Given outcome models $\mu_t(\mathrm{X})$and propensity score $e(\mathbf{X})$, the corresponding pseudo-outcome is defined as shown in Equation (1).

$$\phi(\mathrm{X}, T, Y) = \mu_1(\mathrm{X}) - \mu_0(\mathrm{X}) + \frac{T}{e(\mathrm{X})}(Y - \mu_1(\mathrm{X})) - \frac{1 - T}{1 - e(\mathrm{X})}(Y - \mu_0(\mathrm{X})) \qquad (1)$$

We regress $\phi(\mathrm{X}, T, Y)$on Xusing ridge regression to obtain a regularized CATE estimator. The R-learner adopts a related orthogonalization strategy by residualizing both outcomes and treatment, and estimating $\tau(\mathrm{X})$through minimization of (Equation 2):

$$\sum_i (Y_i - \mu(\mathrm{X}_i) - (T_i - e(\mathrm{X}_i))\,\tau(\mathrm{X}_i))^2 \qquad (2)$$

which, in our linear specification, reduces to a weighted regression problem (Nie & Wager, 2021). In randomized experiments, both DR- and R-learners exploit the known propensity score to improve statistical efficiency and mitigate bias. Importantly, these formulations extend naturally to observational settings in which treatment propensities vary due to selection mechanisms or reporting processes.

L. Cookieless cohort estimators

For cohort-level estimation, we restrict the analysis to aggregate statistics, namely counts $(n_1, y_1)$and (n_0, y_0)corresponding to treated and control units within each cohort. The naive cohort uplift is computed as $y_1/n_1 - y_0/n_0$. Because raw differences are unstable when sample sizes are small, we additionally implement a Bayesian shrinkage estimator based on independent beta–binomial posteriors for each treatment arm. Assuming a Beta$(1,1)$prior, the posterior distribution of the treated conversion rate is Beta$(1 + y_1, 1 + n_1 - y_1)$, with an analogous expression for the control group. Cohort uplift is estimated as the posterior mean difference, and 95% credible intervals are obtained via posterior simulation (4,000 draws).

Table 6. Experimental Setup and Model Hyperparameters

Component	Setting
Train/test split	70% train / 30% test (stratified by treatment)
Random seed	42
Outcome	conversion (binary)
Treatment	Any E-Mail vs No E-Mail
Base outcome model (classification)	LogisticRegression(solver='lbfgs', max_iter=1000)
Outcome model regularization	L2, C=1.0
Ridge stages (X-/DR-/R-learners)	Ridge(alpha=1.0)
Propensity score	Constant e = P(T=1) estimated from training split
Cohort binning (recency)	5 quantile bins computed on training and reused on test
Cohort binning (history)	5 quantile bins computed on training and reused on test
Cohort prior	Beta(1,1) per arm
Cohort posterior MC draws	4000 samples
Bootstrap resamples	200
DP noise	Laplace noise added to cohort counts; scale=1/epsilon
k-anonymity thresholds	k in {5,10,20,50}

This framework reflects the constraints of privacy-preserving aggregate reporting systems, which typically disclose only noisy summary counts. Under such conditions, inference requires explicit regularization and uncertainty quantification, rather than reliance on high-capacity individual-level predictive models.

M. Evaluation metrics

We report (a) the Qini coefficient and (b) the area under the uplift curve (AUUC) to evaluate ranking quality, following standard uplift practice (Radcliffe & Surry, 2011). We also estimate the policy value: the expected conversion rate if we treat only the top p fraction of customers ranked by predicted uplift, estimated on randomized test data using inverse propensity weighting.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 17-38
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Finally, we report train and scoring runtimes to capture computational feasibility. In applied settings, the absolute scale of Qini depends on the sample size and base conversion rate, so we interpret Qini differences primarily in relative terms and in conjunction with downstream policy value lifts and their uncertainty intervals (Table 7–Table 9).

a. Ranking metrics (Qini/AUUC)

Let $i$index test customers, and let $\hat{y}_i$denote the predicted uplift score. Customers are ranked in descending order of $\hat{y}_i$, yielding prefixes of size k. For each prefix, Radcliffe's incremental conversions are computed as Equation 3.

$$\mathrm{Inc}(k) = Y_1(k) - Y_0(k) \cdot \frac{N_1(k)}{N_0(k)} \qquad (3)$$

where $N_t(k)$denotes the number of treated $(t = 1)$ or control $(t = 0)$ customers within the prefix, and $Y_t(k)$is the corresponding number of observed conversions. The scaling factor $N_1(k)/N_0(k)$adjusts for imbalance in treatment and control sample sizes. The Qini curve plots $\mathrm{Inc}(k)$against the population share k/n. The Qini coefficient is defined as the area between the model-specific Qini curve and the random-targeting baseline, represented by the line segment from $(0, 0)$to $(1, \mathrm{Inc}(n))$. The area under the uplift curve (AUUC) corresponds to the unnormalized area under the Qini curve. In our implementation, the curve is evaluated at 100 equally spaced quantiles to stabilize numerical computation and to ensure consistency with standard uplift evaluation dashboards.

b. Policy value estimation

A ranking metric is meaningful only to the extent that it improves decision-making. We therefore estimate the conversion rate achieved u, open paren dot, close paren that assigns treatment to the top p fraction of customers based on predicted uplift and withholds treatment from the remainder. For cohort-only estimators, individual customers inherit their cohort-level score, so the policy reduces to $\pi_p(C)$. Because the test data arise from randomized assignment with known propensity $e \approx \mathbb{P}(T = 1)$, the counterfactual policy value can be estimated using inverse propensity scoring (IPS) (Equation 4):

$$V(\pi_p) = \frac{1}{n} \sum_i \left[\frac{\pi_{p,i} T_i Y_i}{e} + \frac{(1 - \pi_{p,i})(1 - T_i) Y_i}{1 - e} \right] \qquad (4)$$

Under randomized assignment with positivity ($0 < e < 1$) and standard consistency and no-interference assumptions, this estimator is unbiased and requires only that the propensity score is known (or consistently estimated). We report $V(\pi_p)$for $p = 0.2$and $p = 0.5$in Table 7, and present the full policy value curve as a function of $p$in Figure 4.

IV. RESULT

A. Main experimental results

Table 7 reports the primary comparison across all models on the held-out test set. Figure 3 visualizes the Qini curves, and Figure 4 shows the policy value curves over targeting fractions.

Table 7. Main Model Comparison on the Hillstrom Test Split (Qini, AUUC, and Policy Values)

model	qini	auuc	policy_value_20pct	policy_value_50pct	train_seconds	score_seconds
T-learner (LR)	6.674715	35.597744	0.008363	0.00961	3.178047	0.19722
X-learner (LR+Ridge)	5.999817	34.922845	0.007894	0.00961	5.394361	0.102158
DR-learner (AIPW+Ridge)	5.806734	34.729763	0.007816	0.009688	2.00346	0.002562
R-learner (LR+Ridge)	5.513309	34.436337	0.007973	0.009688	3.40293	0.002526
S-learner (LR+int)	4.72604	33.649069	0.008285	0.009297	2.402423	0.103157
Transformed outcome (Ridge)	4.559913	33.482942	0.007582	0.00961	0.079304	0.099448
Random	2.437518	31.360547	0.007582	0.009298	0	0
Cohort-Bayes (k>=10)	-0.286646	28.636383	0.006488	0.009296	0	0
Zero	-0.407462	28.515566	0.00727	0.008516	0	0
Cohort-raw (diff-in-means)	-2.844135	26.078894	0.006644	0.008672	0	0

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 17-38
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

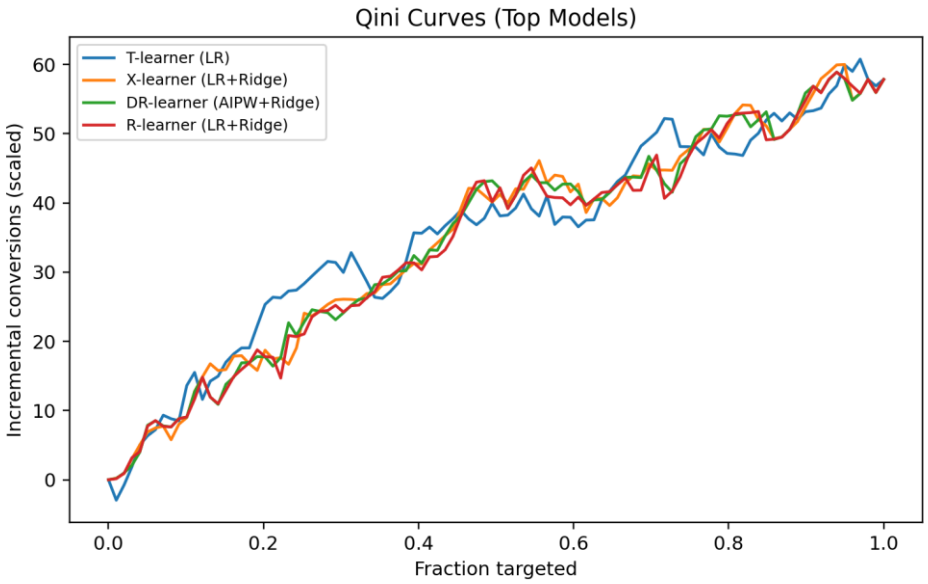


Figure 3. Qini Curves on the Hillstrom Test Split

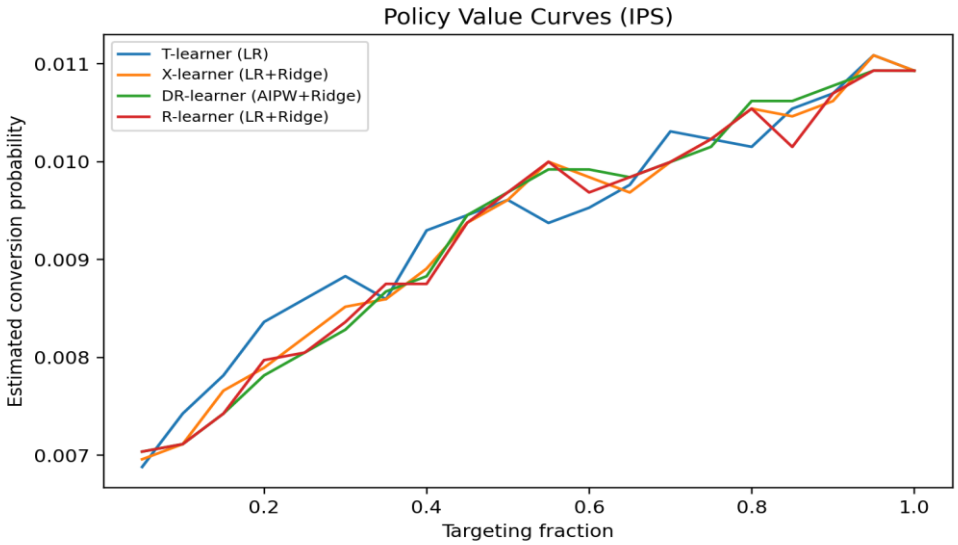


Figure 4. Estimated Policy Conversion Rate as a Function of Targeting Fraction

In our controlled experiment setting, ID-level ITE estimation clearly outperforms cohort-only estimation. The T-learner (logistic regression) achieves the highest Qini (6.675) and the highest AUUC (35.598). DR-learner and X-learner perform comparably (Qini 5.807 and 6.000, respectively) and provide strong policy values. The cohort-only estimators perform close to (and sometimes worse than) random ranking, reflecting information loss from aggregation and the variance of small cohorts. Bayesian shrinkage improves over raw cohort differences in both Qini and AUUC, highlighting the value of regularization when only aggregates are available.

Interpreting policy values (Sun et al., 2023), Table 7 shows that, beyond ranking metrics, the best-performing models also increase the estimated conversion rate under targeted treatment. For example, targeting the top 20% of customers by predicted uplift yields an estimated policy conversion rate of 0.00836 for the T-learner and 0.00789 for the X-learner, compared with 0.00573 when nobody is treated (the No E-Mail conversion rate in Table 3). For context, the ~0.87 Qini gap between the T-learner and DR-learner corresponds to an absolute difference of about 0.00055 in policy conversion rate at p=0.2 in this dataset (Table 7). The absolute gains are modest because baseline conversion rates are low; however, in large-scale advertising and CRM systems, even a 0.0005–0.0010 absolute improvement can translate into meaningful incremental revenue when applied to millions of users. Importantly, these policy values are evaluated out-of-sample on randomized data with IPS, so they reflect causal, not merely correlational, improvements.

Why do some meta-learners perform better? The strong performance of the T-learner and the X/DR learners is consistent with the structure of the Hillstrom data. The experiment randomizes treatment but allows the outcome model to differ across arms, which favors approaches that model treated and control responses separately or that explicitly correct residual bias through propensity weighting. By contrast, the S-learner uses a single parametric model with interactions; with linear features and limited interaction capacity, it can underfit complex heterogeneity. Transformed-outcome regression is unbiased in expectation, but it can be highly variable because it reduces the problem to a single noisy regression target. These observations align with prior uplift evaluations that emphasize the benefits of flexible modeling and robust estimation (Devriendt et al., 2018; Gutierrez & Gérardy, 2017; Künzel et al., 2019).

Table 8. Training and Scoring Runtime (Seconds) by Model

model	train_seconds	score_seconds
T-learner (LR)	3.178047	0.19722
X-learner (LR+Ridge)	5.394361	0.102158
DR-learner (AIPW+Ridge)	2.00346	0.002562
R-learner (LR+Ridge)	3.40293	0.002526
S-learner (LR+int)	2.402423	0.103157
Transformed outcome (Ridge)	0.079304	0.099448
Random	0	0
Cohort-Bayes (k>=10)	0	0
Zero	0	0
Cohort-raw (diff-in-means)	0	0

B. Uncertainty quantification

We quantify the variability of ranking and policy metrics by bootstrapping the test set 200 times. Table 9 summarizes bootstrap means and 95% percentile intervals for representative models. Because Qini and AUUC are high-variance, non-linear functionals of the induced ranking, their bootstrap intervals can be wide and may cross zero; we therefore caution against over-interpreting small Qini differences that fall within these intervals.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 17-38
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 9. Bootstrap Uncertainty for Selected Models (200 Resamples)

model	qini_mean	qini_ci_low	qini_ci_high	v20_mean	v20_ci_low	v20_ci_high
X-learner (LR+Ridge)	5.811872	-3.865685	15.335766	0.007684	0.006055	0.009343
DR-learner (AIPW+Ridge)	5.609519	-4.154505	15.182892	0.007654	0.006014	0.009507
Cohort-Bayes (k>=10)	0.307838	-10.987316	10.86788	0.006405	0.004764	0.008209

C. Privacy robustness experiments

We evaluate how cookieless cohort estimation changes as privacy constraints intensify. Table 10 and Figure 5 vary the minimum cell size k, while Table 11 and Figure 6 vary the differential privacy budget ε.

Table 10. K-Anonymity Robustness for Cohort Bayesian Estimation

k_min	qini	auuc
5	-3.376921	25.546108
10	-0.286646	28.636383
20	-1.997854	26.925175
50	-4.222149	24.700879

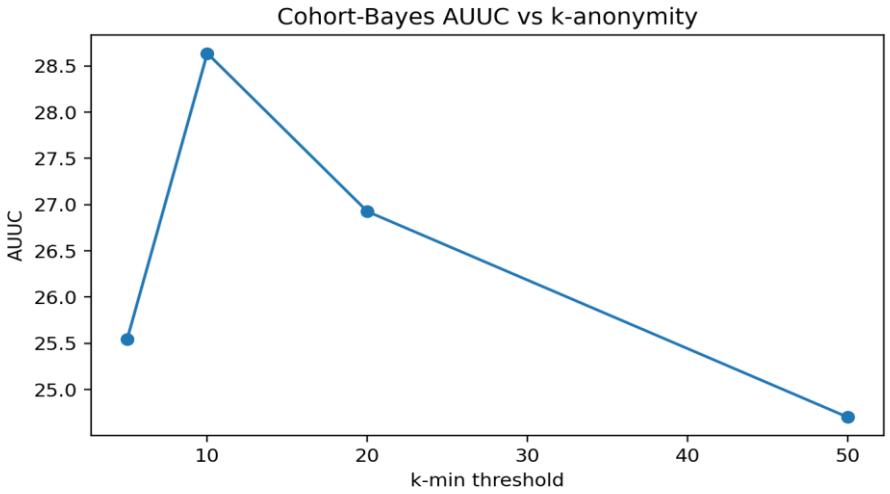


Figure 5. AUUC versus K-Anonymity Threshold (Cohort Bayesian Method)

Table 11. Differential Privacy Robustness for Cohort Bayesian Estimation

epsilon	qini	auuc
0.5	-4.794921	24.128107
1	-7.252234	21.670795
2	-7.336603	21.586426
4	-1.455195	27.467833
8	-0.054437	28.868591

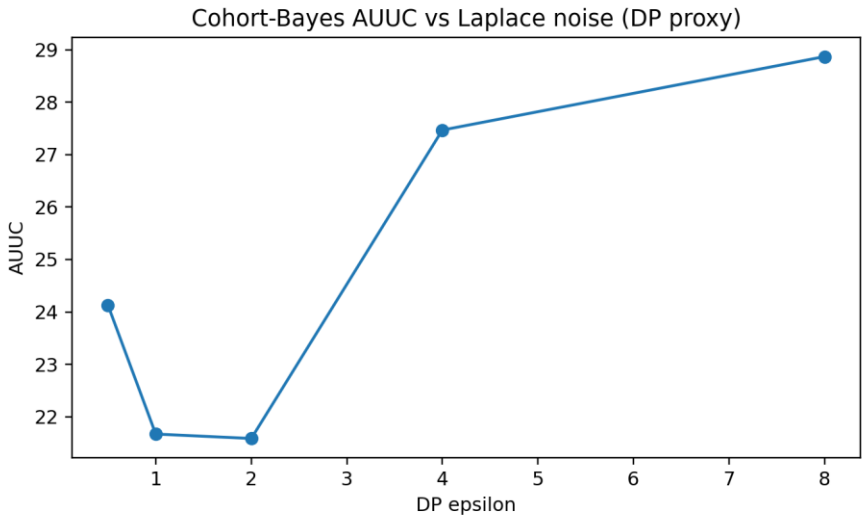
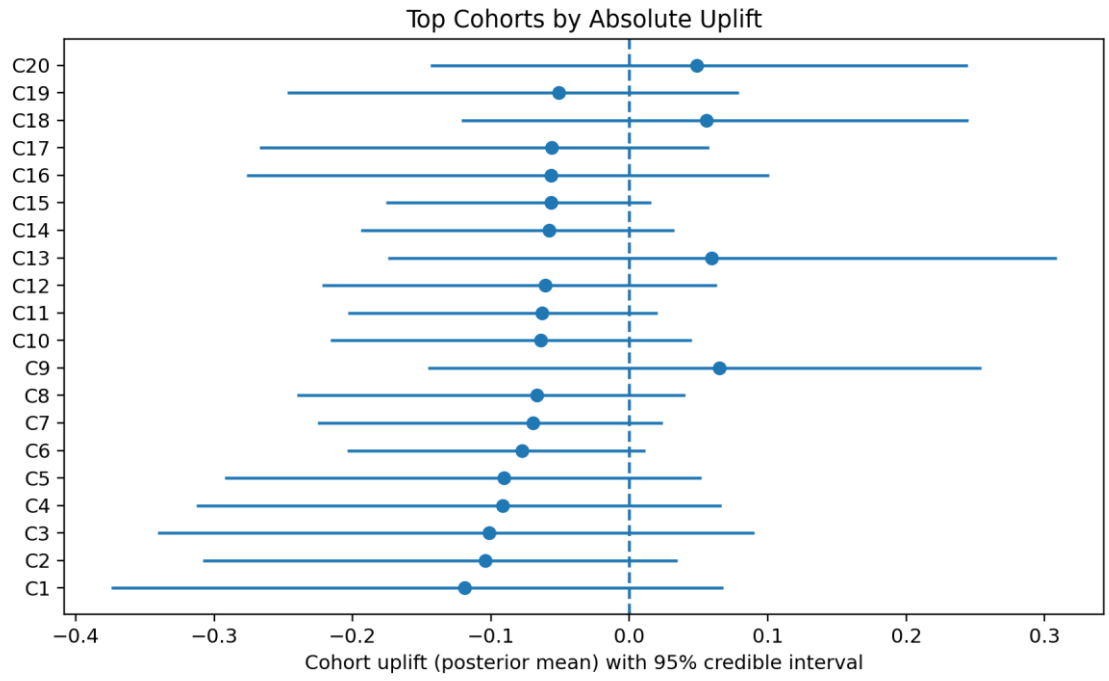


Figure 6. AUUC versus Privacy Budget ε (Cohort Bayesian Method)

The robustness results demonstrate an explicit accuracy–privacy trade-off. Increasing k suppresses more cohorts, reducing targeting resolution and eventually degrading AUUC. Adding stronger DP noise (smaller ε) also reduces AUUC; as ε increases, the noise decreases and performance approaches the baseline of the non-noised cohort.

D. Cohort-level uncertainty intervals

Beyond point estimates, the cohort Bayesian model produces posterior credible intervals for the uplift (difference in conversion rates). Figure 7 shows the top cohorts (by absolute estimated uplift) and their 95% posterior intervals.



Journal of Technology Informatics and Engineering (JTIE)
Vol. 5 No. 1 April 2026
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 17-38
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Figure 7. Posterior Uplift Intervals for Top Cohorts (Training Aggregates)

Table 12. Empirical Coverage and Average Width of 95% Cohort Uplift Intervals

n_cohorts_evaluated	coverage_95	avg_interval_width
231	0.891775	0.1287

E. Interval validation

To validate whether the reported cohort intervals are calibrated, we compare each cohort's training-posterior interval with an independent estimate of uplift computed on the test split. Across 231 evaluable cohorts, the nominal 95% interval covers the test uplift 0.892 of the time, with an average interval width of 0.129 (Table 12). This empirical coverage is below 0.95, indicating under-coverage (intervals that are too narrow), consistent with the fact that the simple beta–binomial model assumes independent Bernoulli outcomes and does not account for cohort definition selection, cohort-level overdispersion, or privacy noise. For monitoring dashboards, these intervals may still be useful as approximate uncertainty bands. However, for decision-making or automated gating one should consider interval-widening strategies (e.g., more conservative or hierarchical/overdispersed models, posterior predictive checks, or empirical calibration/inflation of interval widths to achieve nominal coverage).

V. CONCLUSION AND RECOMMENDATION

This paper studied incrementality estimation in cookieless settings using uplift/ITE models under privacy constraints. Using the Hillstrom randomized e-mail experiment, we conducted full experimental evaluations of multiple ITE estimators and cohort-only cookieless estimators. The randomized design supports clear identification of incrementality: sending an e-mail campaign increases conversion by 0.00495 (95% CI [0.00355, 0.00636]).

Among ID-level methods, meta-learners provide strong ranking quality, with the T-learner (logistic regression) achieving the best Qini and AUUC. Doubly robust learning performs competitively and scores extremely quickly, making it a strong default for production uplift scoring when model monitoring is in place. Cohort-only estimation—representing a cookieless measurement regime—shows a weaker signal and is sensitive to k-anonymity suppression and DP noise, but Bayesian shrinkage and explicit uncertainty reporting mitigate instability.

Based on the results, we make three recommendations for practitioners operating in cookieless environments. First, prioritize experimental design and measurement infrastructure that preserves randomization whenever possible; without randomization, privacy-driven aggregation can substantially increase the risk of bias. Second, favor robust ITE estimators (e.g., doubly robust learners) and report uncertainty, not only point estimates. Bootstrap intervals for policy value and

Bayesian intervals for cohort lift provide complementary views of uncertainty and can be used as monitoring signals. Third, treat privacy parameters as first-class knobs in the measurement system: choose k thresholds and ε budgets based on explicit accuracy targets, and continuously validate downstream decisions against uncertainty-aware metrics.

REFERENCES

Apple. (2021). *Take Advantage of New Advertising Attribution Technologies*. *Apple Developer News*. https://developer.apple.com/news/?id=wajvzt18

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized Random Forests. *Annals of Statistics*, *47*(2), 1148–1178. https://doi.org/10.1214/18-aos1709

Bang, H., & Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, *21*(1), 1–68. https://doi.org/10.1111/ectj.12097

Devriendt, F., Moldovan, D., Verbeke, W., & Baesens, B. (2018). A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone toward the Development of Prescriptive Analytics. *Big Data*, 6(1), 13–41. https://doi.org/10.1089/big.2017.0104

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. *Theory of Cryptography Conference (TCC 2006), Lecture Notes in Computer Science, 3876*, 265–284. https://doi.org/10.1007/11681878_14

Dwork, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3), 211–407. https://doi.org/10.1561/0400000042

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. *Monographs on statistics and applied probability*, *57*(1), 1-436. https://doi.org/10.1201/9780429246593

Google. (2025). *Overview of Attribution Reporting API. Privacy Sandbox*. https://privacysandbox.google.com/private-advertising/attribution-reporting

Gutierrez, P., & Gérardy, J.-Y. (2017). Causal Inference and Uplift Modeling: A Review of the Literature. In *Proceedings of the 3rd International Conference on Predictive Applications and APIs, 67*, 1–13. https://proceedings.mlr.press/v67/gutierrez17a.html

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What If Chapman & Hall/CRC*. https://miguelhernan.org/whatifbook

Hillstrom, K. (2008, March 20). *The MineThatData E-Mail Analytics and Data Mining Challenge*. MineThatData. https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

Hanqi Zhang. (2023). DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models. *Journal of Advanced Computing Systems*, *3*(7), 24-40. https://doi.org/10.69987/jacs.2023.30703

Hanqi Zhang. (2024). Risk-Aware Budget-Constrained Auto-Bidding under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework. *Journal of Advanced Computing Systems*, *4*(6), 30-47. https://doi.org/10.69987/jacs.2024.40603

Hanqi Zhang. (2025). Counterfactual Learning-to-Rank for Ads: Off-Policy Evaluation on the Open Bandit Dataset. *Journal of Advanced Computing Systems*, *5*(12), 1-11. https://doi.org/10.69987/jacs.2025.51201

Jamaludin, H., Achlison, U., & Rokhman, N. (2024). Enhancing AI Model Accuracy and Scalability Through Big Data and Cloud Computing. Journal of Technology Informatics and Engineering, *3*(3), 296–307. https://doi.org/10.51903/jtie.v3i3.203

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press. https://doi.org/10.1017/cbo9781139025751

Jubin Zhang. (2025). Graph-based Knowledge Tracing for Personalized MOOC Path Recommendation. Journal of Advanced Computing Systems, *5*(11), 1-15. https://doi.org/10.69987/jacs.2025.51101

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled Experiments on the Web: Survey and Practical Guide. *Data Mining and Knowledge Discovery*, *18*(1), 140–181. https://doi.org/10.1007/s10618-008-0114-1

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Lo, V. S. Y. (2002). The True Lift Model: A Novel Data Mining Approach to Response Modeling in Database Marketing. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 481–486. ttps://doi.org/10.1145/772862.772872

Nie, X., & Wager, S. (2021). Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *Biometrika*, *108*(2), 299–319. https://doi.org/10.1093/biomet/asaa076

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press. https://www.cambridge.org/9780521895606

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-Learn: Machine Learning in Python.

Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Radcliffe, N. J. (2007). Using Control Groups to Target on Predicted Lift: Building and Assessing Uplift Models. *Direct Marketing Analytics Journal, 1*, 14–21, 1, 14–21. https://doi.org/10.1007/s10796-022-10283-4

Radcliffe, N. J., & Surry, P. D. (2011). Quality Measures for Uplift Models. *Technical report*. https://www.stochasticsolutions.com/pdf/kdd2011late.pdf

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rzepakowski, P., & Jaroszewicz, S. (2012). Decision Trees for Uplift Modeling with Single and Multiple Treatments. *Knowledge and Information Systems*, *32*(2), 303–327. https://doi.org/10.1007/s10115-011-0434-0

Shirakawa, T., Li, Y., Wu, Y., Qiu, S., Li, Y., Zhao, M., Iso, H., & van der Laan, M. (2024). Longitudinal Targeted Minimum Loss-Based Estimation with Temporal-Difference Heterogeneous Transformer. *Proceedings of machine learning research*, *235*, 45097. https://pmc.ncbi.nlm.nih.gov/articles/pmc12681028/

Sklift. (2021). fetch_hillstrom: MineThatData E-Mail Analytics and Data Mining Challenge Dataset (Copy). https://www.uplift-modeling.com/en/v0.3.1/api/datasets/fetch_hillstrom.html

Sweeney, L. (2002). k-anonymity: A Model for Protecting Privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 557–570. https://doi.org/10.1142/s0218488502001648

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Shirakawa, T., Li, Y., Wu, Y., Qiu, S., Li, Y., Zhao, M., Iso, H., & Van der Laan, M. (2024). Longitudinal Targeted Minimum Loss-Based Estimation with Temporal-Difference Heterogeneous Transformer. In *Proceedings of the 41st International Conference on Machine Learning*, *235*, 45097. https://pmc.ncbi.nlm.nih.gov/articles/pmc12681028

Xinzhuo Sun, Yifei Lu, & Jing Chen. (2023). Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting. *Journal of Advanced Computing Systems*, *3*(8), 9-24. https://doi.org/10.69987/jacs.2023.30802

Xinzhuo Sun, Jing Chen, Binghua Zhou, & Meng-Ju Kuo. (2024). ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence. *Journal of Advanced Computing Systems*, *4*(7), 50-64. https://doi.org/10.69987/jacs.2024.40705