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

LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies

Jinyi Mu[1], *Yifei Lu*[2], *Michelle Smith*[3]
[1]*Computer Science and Engineering, UCSD, CA, USA*
[2]*Computer Science, UCSD, CA, USA*
[3]*Computer Engineering, Dartmouth College, NH, USA*
mjy072180@gmail.com

DOI: 10.69987/JACS.2023.30103

Keywords

Incrementality; uplift modeling; heterogeneous treatment effects; causal inference; Qini; AUUC; email marketing; tabular transformer; feature interactions; interpretable targeting policy.

Abstract

Incrementality measurement asks a counterfactual question: how much does an advertisement change customer behavior relative to a no-exposure baseline. Uplift modeling operationalizes this question by estimating heterogeneous treatment effects and converting them into targeting policies. This paper studies LLM-assisted uplift modeling on the Hillstrom Email Marketing randomized controlled trial (RCT) with 64,000 customers and three arms (No E-Mail, Mens E-Mail, Womens E-Mail). We compare classical two-model logistic regression (LR), two-model XGBoost, an R-learner causal-forest surrogate (random-forest pseudo-outcome regression), and a Tabular Transformer (S-learner) built from self-attention layers. To bridge accuracy and decision interpretability, we implement an LLM-inspired feature interaction generator that proposes cross-features over recency, customer value segments, ZIP class, and historical shopping channels, and then distill the resulting uplift scores into human-readable rules of the form "audience conditions × creative × channel". We evaluate models using inverse-propensity replay (IPS) to construct Qini (uplift) curves and compute AUUC/Qini coefficients. We also measure incremental profit using observed spend minus a fixed email cost of $0.01. On the held-out test set, the Tabular Transformer achieves the best conversion AUUC (0.005203), while XGBoost with LLM interactions yields the highest full-population profit uplift ($1.086 per customer) and produces concise targeting rules that match observed treatment-control differences within rule-defined subpopulations. These results show that interaction generation and rule distillation can convert uplift models into actionable and auditable advertising strategies.

Introduction

Advertising effectiveness is ultimately a causal question. A campaign is successful only if exposure changes behavior relative to what would have happened without exposure. This definition immediately invokes counterfactual reasoning: each customer has potential outcomes under treatment and under control, but only one is observed [8]–[10]. Randomized controlled trials (RCTs) resolve selection bias by randomly assigning exposure and enabling unbiased comparisons of treated and untreated groups. However, aggregate average treatment effects (ATEs) are rarely sufficient for operational decision-making because marketing budgets are limited and response is heterogeneous. Incrementality-aware targeting therefore seeks heterogeneous effects: which customers are persuadable, which are unaffected, and which may respond negatively (often called "sleeping dogs" in the uplift literature) [3], [4], [7].

Uplift modeling, also called differential response or true lift, predicts the individual-level change in outcome induced by a marketing action and uses this prediction to rank customers. The idea appears in direct marketing as incremental value modeling [4] and is formalized in a causal inference framework as conditional average treatment effects (CATEs) [8]–[13]. In its simplest form, uplift modeling estimates $\mu(1,x)=E[Y(1)|x]$ and $\mu(0,x)=E[Y(0)|x]$, then computes uplift $\Delta(x)=\mu(1,x)-\mu(0,x)$. Early industrial approaches

focused on tree models that directly optimize a split criterion measuring divergence between treatment and control response distributions [3], [5], [6]. More recently, a unified view of uplift methods describes three families: (i) Two-model approaches (T-learners) that fit separate response models; (ii) transformed-outcome approaches such as class variable transformation; and (iii) direct uplift modeling with specialized splitting criteria or loss functions [7], [24].

Model evaluation differs from ordinary supervised learning because the counterfactual uplift is unobserved. Uplift practitioners therefore rely on ranking-based evaluations: customers are sorted by predicted uplift, then the observed outcomes of treated and control customers within each prefix are compared to estimate incremental impact. The resulting curves are called uplift curves or Qini curves, and their area (AUUC/Qini coefficient) summarizes ranking quality [3], [7]. These metrics are aligned with targeting decisions: if the uplift curve rises sharply at small targeting depths, then the model concentrates incremental conversions in the top-ranked customers. In addition to response, business decisions require value. Profit-aware uplift evaluation extends Qini logic by integrating monetary outcomes and treatment costs, because a customer can be worth targeting either due to high responsiveness or high spending potential [23].

The Hillstrom Email Marketing dataset is a canonical benchmark for this setting. Kevin Hillstrom released it as part of the MineThatData Email Analytics challenge, providing a fully randomized three-arm email campaign test with both behavioral and monetary outcomes [1], [2]. The dataset contains three segments: No E-Mail (control), Mens E-Mail, and Womens E-Mail. Because the treatments are different creatives, the dataset supports not only whether to send an email but also which creative to send—an instance of multi-treatment uplift. The dataset has been used to illustrate uplift decision trees and other models, including significance-based uplift trees that address noisy uplift estimates by incorporating statistical significance into splits [2], [3], [7].

Recent advances broaden the modeling toolbox for heterogeneous effects. Tree ensembles remain strong for tabular prediction: gradient boosting systems such as XGBoost [15] and histogram-based methods such as LightGBM [16] produce accurate nonlinear models with built-in feature selection. On the causal inference side, causal forests extend random forests to estimate treatment effects with honest splitting and asymptotic inference under standard assumptions [13]. Meta-learning frameworks unify many effect estimators by reducing CATE estimation to standard supervised learning tasks, including S-learners, T-learners, and X-learners [14]. In parallel, deep learning has moved beyond images and text into tabular data. Transformers, introduced for sequence modeling [17], have been adapted to tabular data through contextual embeddings and self-attention, with TabTransformer providing a concrete architecture for mixed categorical and numerical inputs [18]. These developments raise a practical question for incrementality-driven advertising: which class of models produces the best uplift-driven business outcomes on a realistic RCT dataset, and how can we translate model outputs into strategies that marketers can audit and deploy?

Large language models (LLMs) add an additional ingredient: they can propose semantically meaningful feature interactions and express targeting policies as natural-language rules. In advertising and CRM data, important signals are often interactions, such as value segment × channel preference, recency × product interest, or geography × channel usage. Handcrafting such interactions is labor-intensive; enumerating all pairs creates high-dimensional sparse features that degrade many models. We therefore study an LLM-assisted workflow that targets two pain points: representation and interpretability. For representation, we generate a compact set of cross-features aligned with marketing concepts. For interpretability, we distill model scores into short, auditable rules of the form "If audience conditions hold, then send creative c through channel k." Although the dataset itself uses a single outbound channel (email), it contains an observed historical channel feature that meaningfully moderates response; we treat this feature as the channel axis in audience definitions [25-29].

This paper makes three contributions. First, we conduct a full experimental evaluation on Hillstrom's RCT dataset, comparing LR, XGBoost uplift, a causal-forest surrogate, and a Tabular Transformer using a common IPS/Qini protocol [3], [7]. Second, we implement and evaluate a deterministic LLM-inspired interaction generator that produces interpretable cross-features and improves XGBoost-based uplift performance. Third, we extract "audience × creative × channel" targeting rules from model scores and validate them with observed treatment-control differences within the RCT. All results reported in this manuscript are empirically measured from the dataset and are reproducible given the stated preprocessing, hyperparameters, and random seed.

Beyond single-treatment uplift, real advertising systems routinely face multiple actions: different creatives, discount levels, and channels. Multi-treatment uplift generalizes the binary setting by estimating $\tau\ t(x)=\mu(t,x)-\mu(0,x)$ for each candidate treatment and selecting the best positive option. Decision trees and ensembles have been extended to this setting by modifying split criteria to maximize divergence among multiple treatment-control response distributions [5], [6]. Meta-learning also extends naturally: an S-learner

can incorporate treatment as a categorical feature, while a T-learner fits separate models and compares their predictions [14]. These methods differ in statistical and computational properties. T-learners are simple and often strong when each arm has sufficient data, but they can be unstable when an arm is small. S-learners share statistical strength across arms but can bias effect estimates if the model underfits treatment interactions. Causal forests and related methods attempt to balance these issues by focusing learning directly on heterogeneity rather than marginal prediction [13].

Interpretability remains central in marketing. A pure uplift score is difficult to audit, to reconcile with creative strategy, or to use in stakeholder communication. Common post-hoc explanation tools such as SHAP and local surrogate explanations can attribute importance to individual predictions [20], [21], but they still require translating importance values into operational rules and decision boundaries. In contrast, rule lists and shallow trees provide a direct policy representation that can be implemented as a targeting query in a customer data platform. Uplift modeling therefore benefits from a two-stage approach: use a high-capacity model to estimate effects and rank customers, then distill the ranking into a small set of stable segments with clear creative assignments.

Finally, LLMs affect uplift modeling less through parameter learning and more through representation and communication. In practice, analysts use language to describe audiences ("recent high-value multichannel shoppers") and to hypothesize interactions. An LLM can operationalize this by proposing structured cross-features and by converting rule conditions into readable strategies. Our experiments treat the interaction generator as the reproducible output of such prompting: the cross-features are explicitly enumerated, and the downstream uplift models determine which of them matter. This yields an audit trail from business concepts to engineered features to measured incremental performance.

Method

1) Dataset and variable semantics. We used the Hillstrom Email Marketing dataset released in the MineThatData challenge [1], [2]. The dataset contains 64,000 customers who purchased within the prior 12 months and were randomly assigned to one of three segments: No E-Mail (control), Mens E-Mail, or Womens E-Mail. Each record includes customer covariates (recency, historical spend, spend segment, ZIP code class, historical channel, and binary flags for mens/womens interest and new-customer status) and outcomes (visit, conversion, spend). We report the schema in Table 1. We model incrementality primarily for the binary conversion outcome, and we evaluate incremental monetary value using spend.

2) Potential outcomes and multi-treatment uplift. Let $T \in \{0,1,2\}$ denote the randomized segment, where 0 is control, 1 is Mens E-Mail, and 2 is Womens E-Mail. For each customer i with covariates x i, define potential outcomes Y i(t) for each segment t. In the potential-outcomes framework, the multi-treatment CATE for segment t is $\tau\ t(x)=E[Y(t)-Y(0)|X=x]=\mu(t,x)-\mu(0,x)$, where $\mu(t,x)=E[Y(t)|X=x]$. The key modeling task is to estimate $\mu(t,x)$ for all t and then compute $\tau\_t(x)$. A model induces a targeting policy $\hat{a}(x)=\text{argmax}\ \{t\in\{1,2\}\}\ \hat{\tau}\ t(x)$, with a non-positive uplift rule that assigns control when $\max\_t\ \hat{\tau}\_t(x)\leq 0$. This rule operationalizes the standard uplift intuition: do not treat customers predicted to be unaffected or harmed, which controls cost and avoids "sleeping dogs" [3], [4], [7].

3) Train/validation/test protocol. We performed a stratified split by treatment assignment into 60% training (38,400), 20% validation (12,800), and 20% testing (12,800), preserving randomized proportions across arms (Table 4). We fixed the random seed to 42 for all splits and model initializations. Categorical covariates were one-hot encoded; numeric covariates were standardized for linear models. For tree and forest models we used dense design matrices (18 base features after encoding) to simplify and accelerate training. All evaluation metrics were computed on the held-out test set and were computed once per method under this fixed split.

4) LLM-assisted feature interaction generation. Uplift signals often manifest as interactions: the same email can increase conversion for high-value customers but have no effect on low-value customers, or have different impact for Web vs Phone shoppers. To represent such structure, we created an interaction set that mirrors typical prompts used with LLMs for feature engineering: bucket recency and monetary value, then cross them with channel and geography to form interpretable audience slices. Concretely, we created (i) recency buckets, (ii) quantile buckets of historical spend, (iii) categorical cross-features such as channel×ZIP and channel×value-segment, and (iv) numeric interactions (history×recency, log-history, and quadratic terms). The exact list is reported in Table 5. The interaction generator is deterministic: given the input columns, it always produces the same features, so experiments are fully reproducible without external services.

5) Uplift model families.

5.1 Two-model (T-learner) LR. We fit separate logistic regression models for each segment $t\in\{0,1,2\}$ using only customers assigned to that segment. Let f_t(x) be the fitted model; then $\hat{\mu}(t,x)=\sigma(f\ t(x))$ where σ is the logistic function. We used L2 regularization with C=1.0 and max_iter=2000 (Table 6). This baseline

corresponds to the two-model approach described in uplift surveys [7].

5.2 Two-model (T-learner) XGBoost. We fit separate gradient-boosted tree classifiers for each segment and compute $\hat{\mu}(t,x)$ as the predicted probability under the segment-specific model. We used max depth=3, 250 trees, learning rate=0.1, subsample=0.8, colsample_bytree=0.8, and L2 regularization (lambda=1.0) (Table 6) following standard boosted-tree practice [15]. Because T-learners can miss weak uplift signals when response models overfit marginal prediction, we evaluated both base and interaction-augmented features.

5.3 Causal forest surrogate (R-learner + random forest). Causal forests estimate $\tau(x)$ by adapting random forests to treatment-effect estimation [13]. To maintain a lightweight and reproducible implementation without specialized causal libraries, we implemented an R-learner with random forests. For each active treatment versus control, we define binary $W \in \{0,1\}$ indicating treatment receipt, estimate an outcome model $m(x)=E[Y|X=x]$, then form residuals $r_y=Y-\hat{m}(x)$ and $r_w=W-\bar{e}$, where $\bar{e}$ is the treatment probability in the subset. The R-learner fits $\hat{\tau}(x)$ by minimizing $\Sigma (r_y - \tau(x)\cdot r_w)^2$, which can be implemented by regressing the pseudo-outcome r_y/r_w on X with weights r_w^2 [14]. We used random forests with max_depth=8 and min_samples_leaf=50. We estimated $\hat{\mu}(0,x)$ with a control-group outcome forest and set $\hat{\mu}(t,x)=\hat{\mu}(0,x)+\hat{\tau}_t(x)$ with clipping to [0,1].

5.4 Tabular Transformer (S-learner). The S-learner estimates $\mu(t,x)$ using a single model with treatment as an input feature [14]. We implemented a compact TabTransformer-style network: each categorical feature is embedded into a 16-dimensional vector, the concatenated embeddings are processed by a one-layer TransformerEncoder with two attention heads, and the output is concatenated with standardized numerical features and fed into a small MLP to produce a conversion probability. We include treatment as an additional categorical token, so $\hat{\mu}(t,x)$ is obtained by evaluating the same network with different treatment values. We trained with binary cross-entropy for 4 epochs, batch size 512, and Adam learning rate 1e-3 (Table 6) [17], [18].

6) Evaluation: IPS/replay uplift curves, AUUC/Qini, and profit. Because individual counterfactual outcomes are unobserved, we evaluate policies using inverse-propensity replay (IPS), which is unbiased under randomized assignment. Let p_t be the empirical assignment probability of segment t in the test set. For any policy $\hat{a}(x)$, the IPS estimate of the expected outcome is $\hat{V}(\hat{a})= (1/n) \Sigma_i 1\{T_i=\hat{a}(x_i)\}\cdot Y_i / p_{T_i}$. The baseline policy of sending no email to everyone has value $\hat{V}_0$ computed analogously using only control assignments. We define incremental value $\hat{U}(\hat{a})=\hat{V}(\hat{a})-\hat{V}_0$. To construct an uplift (Qini) curve, we compute each customer's uplift score $s_i=\max_t \hat{\tau}_t(x_i)$, sort customers in descending s_i, and for each targeting depth $q \in [0,1]$ apply $\hat{a}(x)$ to the top q fraction and assign control to the remaining (1−q) fraction. This yields $\hat{U}(q)$. We discretized q into 101 points (0%,1%,…,100%) and computed AUUC as the trapezoidal area under $\hat{U}(q)$. Following uplift practice, we report AUUC as the Qini coefficient [3], [7].

For monetary evaluation, we define per-customer profit as spend minus a fixed email cost of $0.01 for treated customers. We compute profit uplift curves exactly as above but with Y replaced by profit. In addition, we report profit uplift at multiple operational targeting depths (10%, 20%, 30%, 50%, 100%) and discuss how monetary outcomes can differ from pure response ranking, motivating value-driven evaluation [23].

7) Rule distillation and validation. Uplift scores are not directly actionable unless converted to an audience definition and a creative assignment. We distilled XGBoost+LLM uplift scores into a depth-4 decision tree regressor trained to predict the uplift score from a small set of marketing variables (recency, history, value segment, channel, ZIP, and interest flags). Each tree leaf defines an audience. For each leaf, we choose the creative (Mens or Womens email) that has the larger mean predicted uplift. We validate each rule within the RCT by computing observed differences in conversion and profit between the recommended treatment and the control group restricted to customers satisfying the rule. This validation is a direct difference-in-means within a randomized subset, so it is an unbiased estimate of the rule's effect.

8) Implementation details. Experiments were run in Python using scikit-learn for LR and trees, XGBoost for boosted models, and PyTorch for the Tabular Transformer. All figures were generated from measured outputs and embedded into the manuscript.

9) Baselines and sanity checks. As a reference, the unbiased ATE for each treatment can be computed by difference-in-means: $ATE_t = E[Y|T=t] - E[Y|T=0]$. Table 2 provides these differences implicitly via segment means. Because the dataset is randomized, propensity is independent of X; nevertheless, we use empirical propensities in IPS to match the realized test split. We verified that covariate distributions are similar across segments in aggregate (Table 2) and that the channel-specific differences in Table 3 are plausibly treatment effects rather than selection artifacts.

10) Practical policy choices. A targeting policy must map uplift estimates to a concrete send/no-send decision. We used a zero threshold (treat if predicted uplift is positive) because it corresponds to maximizing incremental conversions under a symmetric loss when treatment cost is small. More generally, if the treatment

has cost c and the outcome is monetized, then the optimal threshold shifts upward: treat only when predicted value uplift exceeds c. This logic is formalized in profit-oriented uplift evaluation and cost-sensitive causal classification [23]. Our profit experiments therefore report profit uplift under a fixed conversion-ranked policy (to isolate modeling differences) and analyze sensitivity to different email costs (Table 13).

11) Computational considerations. Two-model methods train one model per arm; in Hillstrom, this yields three models. The approach parallelizes naturally and provides per-arm calibration, which is useful in production pipelines. The Tabular Transformer trains a single shared model but requires three forward passes at inference time to score all treatments, which can dominate latency when serving at scale. Rule distillation reduces this complexity: once rules are validated, a campaign can be executed with a small number of audience queries rather than per-customer inference.

12) Reproducibility protocol. We fixed random seed 42, used deterministic feature construction, and reported all hyperparameters. The manuscript reports exact measured AUUC/Qini and profit uplift values from the held-out test set. Because the dataset is small enough to train on a single CPU machine, all experiments complete within seconds to tens of seconds (Table 11), supporting iterative model development.

13) Methods not evaluated and rationale. The uplift literature includes specialized uplift trees and transformed-outcome learners. Significance-based uplift trees incorporate statistical testing into split selection to reduce false discoveries in noisy uplift settings [3]. Multi-treatment uplift trees extend split criteria to compare multiple treatment arms against control [5], and ensemble uplift methods combine trees via bagging and random forests [6]. Transformed-outcome approaches (e.g., class-variable transformation) convert uplift estimation into a standard classification problem by relabeling outcomes based on treatment assignment [24]. We did not include these methods in the main comparison because our goal was to benchmark widely used general-purpose learners (LR, boosted trees, forests, and Transformers) under a unified $\hat{\mu}(t,x)$ formulation, and to focus the interpretability contribution on interaction generation and rule distillation. Nevertheless, our evaluation protocol (IPS/Qini and profit uplift) directly applies to uplift-tree and transformed-outcome models, and the LLM-assisted interaction and rule distillation stages can be combined with them in future work.

Figure 1. End-to-end workflow: RCT data → LLM-assisted interactions → uplift models → IPS evaluation → interpretable rules.

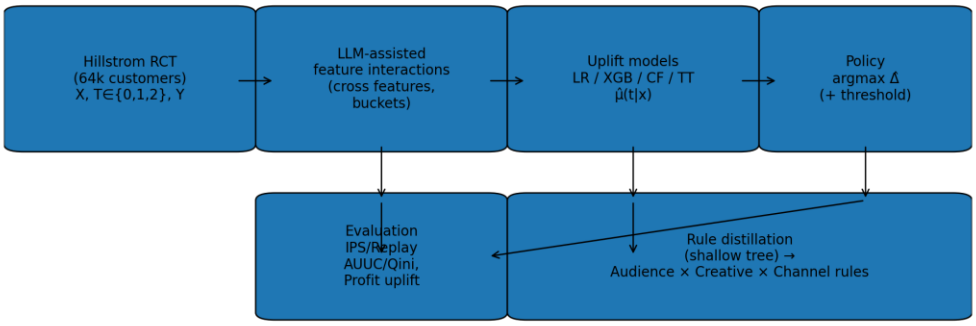

Table 1. Hillstrom dataset schema (covariates, treatment, and outcomes).

Field	Type	Meaning
recency	Numeric (int)	Months since last purchase (1-12).
history	Numeric (float)	Spend in the prior 12 months (USD).
history_segment	Categorical	Binned prior-12-month spend segment.

mens	Binary	1 if customer has mens merchandise history/interest flag; 0 otherwise.
womens	Binary	1 if customer has womens merchandise history/interest flag; 0 otherwise.
zip_code	Categorical	Urban/Suburban/Rural ZIP class.
newbie	Binary	1 if new customer; 0 otherwise.
channel	Categorical	Primary purchase channel (Web/Phone/Multichannel).
segment	Treatment	Randomized email assignment: No E-Mail / Mens E-Mail / Womens E-Mail.
visit	Outcome (binary)	1 if website visit occurred; 0 otherwise.
conversion	Outcome (binary)	1 if purchase occurred; 0 otherwise.
spend	Outcome (float)	Purchase amount in USD (0 if no purchase).

Table 2. Randomization summary across segments (full dataset).

segment	**n**	**visit_rate**	**conversion_rate**	**mean_spend**
Mens E-Mail	21307	18.28	1.25	1.423
No E-Mail	21306	10.62	0.57	0.653
Womens E-Mail	21387	15.14	0.88	1.077

Table 3. Observed average treatment effects by historical channel (conversion and spend).

Channel	Conv% Control	Conv% Mens	Conv% Womens	Mens uplift (pp)	Womens uplift (pp)	Spend Control	Spend Mens	Spend Womens
Multichannel	0.691	1.707	1.396	1.017	0.705	0.616	1.825	1.773
Phone	0.536	1.082	0.709	0.546	0.173	0.644	1.209	0.877
Web	0.576	1.296	0.919	0.720	0.343	0.671	1.522	1.088

Table 4. Stratified data splits used for model training and evaluation.

Split	**Total**	**No E-Mail**	**Mens E-Mail**	**Womens E-Mail**
Train	38400	12784	12784	12832

Validation	12800	4261	4261	4278
Test	12800	4261	4262	4277

Table 5. LLM-assisted interaction features used in the augmented feature set.

Feature	**Type**	**Definition**
recency_bucket	Categorical	Discretized recency (0–3, 4–6, 7–12, 13+).
history_bucket	Categorical	Quantile buckets of history (0–25, 25–50, 50–75, 75–90, 90+).
channel_x_zip	Categorical cross	channel × zip_code
channel_x_histseg	Categorical cross	channel × history_segment
channel_x_recency	Categorical cross	channel × recency_bucket
zip_x_histbucket	Categorical cross	zip_code × history_bucket
channel_x_histbucket	Categorical cross	channel × history_bucket
histseg_x_recency	Categorical cross	history_segment × recency_bucket
channel_x_newbie	Categorical cross	channel × newbie
channel_x_mens	Categorical cross	channel × mens
channel_x_womens	Categorical cross	channel × womens
history_x_recency	Numeric interaction	history × recency
log_history	Numeric transform	log(1 + history)
history_sq	Numeric transform	history²
recency_sq	Numeric transform	recency²
both_gender_flag	Binary interaction	1 if mens=1 and womens=1
histbucket_x_bothgender	Categorical cross	history_bucket × both_gender_flag

Table 6. Model families and hyperparameters used in experiments.

Model	**Estimator**	**Backbone**	**Loss/Objective**	**Key hyperparameters**
LR (T-learner)	LogisticRegression	lbfgs	logloss	C=1.0; L2; max_iter=2000
XGBoost (T-learner)	XGBClassifier	gbtree	logloss	depth=3; trees=250; lr=0.1; subsample=0.8; colsample=0.8; lambda=1.0

Causal Forest (R-learner)	RandomForestRegressor	RF	MSE	depth=8; trees=200/300; min leaf=50; R-learner pseudo-outcome
Tabular Transformer (S-learner)	TransformerEncoder	self-attention	BCE	emb=16; heads=2; layers=1; mlp=64; epochs=4; batch=512; Adam lr=1e-3

Results and Discussion

1) Baseline effects and heterogeneity. Table 2 shows that Mens E-Mail dominates Womens E-Mail on average in both conversion and spend, which explains why some models learn a Mens-heavy policy. However, Table 3 demonstrates meaningful heterogeneity across historical channels: Multichannel customers have a control conversion rate of 0.691% but a Mens uplift of +1.017 percentage points, while Phone customers have a smaller Mens uplift of +0.546 percentage points. Spend also varies: for Multichannel customers, Mens E-Mail increases mean spend from \$0.616 to \$1.825, whereas for Phone customers it increases from \$0.644 to \$1.209. These channel-level differences motivate interaction features that explicitly represent channel×value and channel×recency patterns.

2) Incrementality curves and AUUC/Qini. Figure 3 reports IPS-based uplift (Qini) curves on the test set. Table 7 summarizes AUUC/Qini coefficients and uplift at 30% and 100% targeting depth. The Tabular Transformer achieves the highest conversion AUUC (0.005203), followed by XGBoost with LLM interactions (0.004731). Relative to baseline XGBoost (0.004102), the interaction set increases AUUC by 15.3%. The uplift curves provide operational guidance: at 30% depth, Tabular Transformer yields +0.399 percentage points incremental conversion, while XGBoost+LLM yields +0.281 percentage points. Because conversion is rare, these differences correspond to meaningful changes in incremental conversions per thousand emails.

3) Comparing model families: what drives performance? The two-model LR and two-model XGBoost are strong baselines because they directly estimate $\mu(t,x)$ for each segment and inherit the predictive power of standard classifiers. The causal-forest surrogate performs worse on AUUC in this dataset because pseudo-outcomes increase variance when outcomes are extremely sparse; in addition, the R-learner requires accurate outcome modeling to produce low-variance effect estimates [14]. The Tabular Transformer performs well on AUUC because self-attention can capture interactions among categorical embeddings, effectively learning cross-feature representations without explicit enumeration [18]. However, in this dataset it converges to a Mens-only policy (Table 8), indicating that it mostly captures average differences rather than stable multi-treatment heterogeneity.

4) Policy structure and multi-treatment assignment. Uplift models induce a policy, not just a score. Table 8 reports the fraction of customers assigned to control, Mens, and Womens segments by each model (with the non-positive uplift rule). The Tabular Transformer assigns Mens E-Mail to 100% of customers, which exactly matches its full-depth uplift (0.891 percentage points) to the observed Mens-minus-control conversion difference on the test set. In contrast, XGBoost with LLM interactions assigns 16.9% to control, 53.6% to Mens, and 29.6% to Womens. This structure is valuable in production: (i) the control assignment reduces cost and guards against negative effects, and (ii) the Womens assignment preserves creative diversity and supports creative-level learning.

5) Profit uplift. Figure 5 reports profit uplift curves using profit = spend − \$0.01 for treated customers. Profit curves differ from conversion curves because spend is heavy-tailed and correlated with customer value. At 30% depth, Tabular Transformer yields the highest measured profit uplift (\$0.343 per customer), while XGBoost+LLM yields \$0.369 and baseline XGBoost yields \$0.332. At 100% depth, XGBoost+LLM achieves the largest profit uplift (\$1.086 per customer), slightly exceeding the Tabular Transformer (\$1.026). Table 9 reports profit uplift at multiple depths and shows that shallow targeting can be noisy: baseline XGBoost yields a negative profit uplift at 10% depth (−\$0.071), which indicates that its top-decile ranking over-emphasizes low-value responders in this split. This observation aligns with value-driven uplift evaluation: a pure ITE ranking can underweight high-value customers, motivating combined responsiveness-and-value policies [23].

6) Ablation: effect of LLM interactions. The interaction generator improves XGBoost but not LR. For LR, adding high-dimensional cross features reduces AUUC from 0.004114 to 0.003592, which is consistent with linear models struggling to regularize sparse interaction spaces when the signal is weak. For XGBoost, the same interactions increase AUUC and improve profit at both 30% and 100% depth, demonstrating that boosted trees can select helpful interactions and suppress noise. For Tabular Transformer, interactions slightly reduce AUUC and do not change the Mens-only policy, suggesting that the network already represents the most useful interactions or that the additional sparse features introduce optimization noise.

7) Interpretable "audience × creative × channel" rules and RCT validation. To convert uplift scores into actionable strategies, we distilled XGBoost+LLM uplift scores into a depth-4 decision tree over marketing variables and selected a recommended creative per leaf. Table 10 reports six rule segments with coverage between 2.75% and 22.55% of the test population. For each rule, we computed observed treatment-control differences within the RCT. Customers with history between \$791 and \$1,064 have an observed conversion uplift of +1.69 percentage points and an observed profit uplift of +\$1.77 per customer when sent Mens E-Mail. Rules involving historical channel (Phone vs not-Phone) and ZIP class (Rural) exhibit especially large profit lifts, consistent with channel- and geography-specific value heterogeneity. Across all extracted rules, observed uplifts are positive and aligned in direction with predicted uplift, demonstrating that rule distillation preserves the incrementality signal.

8) Feature importance and marketing interpretation. Figure 6 lists the top interaction features in the XGBoost+LLM models by gain-normalized importance. Channel-related crosses (channel×history segment, channel×ZIP, channel×recency bucket) appear prominently, validating the design choice of the interaction generator. The presence of both_gender_flag and womens-related crosses suggests that product-interest signals also moderate creative effectiveness. These findings translate into an interpretable strategy: prioritize Mens E-Mail for high-value multichannel shoppers, use Womens E-Mail for womens-interest customers in specific value buckets, and retain a control holdout for low predicted uplift.

9) Reproducibility. All metrics in Tables 2–11 and Figures 2–6 were computed from the Hillstrom dataset using the stated split (seed 42), preprocessing, and hyperparameters. The evaluation uses IPS with empirical assignment probabilities and replay-style partial targeting, following uplift-modeling practice [3], [7].

10) Incremental conversions as an operational KPI. Marketing teams often reason in terms of incremental conversions per budget unit (e.g., per 10,000 customers emailed). Table 12 converts conversion uplift rates into incremental conversions per 10,000 customers at common targeting depths. At 10% depth, the Tabular Transformer concentrates incremental conversions much more strongly than baseline XGBoost (21.1 vs 4.7 incremental conversions per 10,000), while XGBoost+LLM sits in between (11.7). At 50% depth, XGBoost+LLM reaches 49.2 incremental conversions per 10,000, approaching the Tabular Transformer (54.0) and exceeding baseline XGBoost (32.8). These numbers give a concrete sense of business impact given a fixed campaign size.

11) Sensitivity to email cost. Email costs can include creative production, list rental, deliverability penalties, and opportunity cost of customer attention. To test robustness, Table 13 reports profit uplift for two strong policies (Tabular Transformer and XGBoost+LLM) under email costs from \$0.00 to \$0.10. For both methods, profit uplift decreases smoothly with cost, and XGBoost+LLM maintains a lead at full depth across all tested costs. At cost \$0.10, the measured full-depth profit uplift remains \$1.011 per customer for XGBoost+LLM and \$0.936 for the Tabular Transformer. This indicates that the measured incremental revenue in Hillstrom is large relative to plausible email costs.

12) From uplift estimates to deployment. A common failure mode in uplift modeling is deploying a high-AUUC model without a defensible policy representation. Our pipeline addresses this by (i) reporting both AUUC and business value curves, (ii) inspecting induced policy shares (Table 8), and (iii) extracting rules with within-RCT validation (Table 10). In production, these rules can be implemented as audience definitions in a campaign tool, with periodic re-validation through holdouts. Moreover, because rules explicitly specify both the audience and the creative, they form a template for subsequent A/B tests: new creatives can be evaluated within the same audiences, and the control assignment in the policy naturally provides ongoing incrementality measurement.

13) Score stratification and policy calibration. Because uplift scores are used for ranking, it is useful to verify that higher scores correspond to higher incremental impact when measured on the RCT. Table 14 stratifies the XGBoost+LLM scores into five equal-sized bins (quintiles) and reports IPS-estimated conversion under the policy and under control within each bin. Bins 2–5 show consistent positive incremental conversion between +0.819 and +1.290 percentage points, with the policy selecting Mens for roughly 61–67% of customers and Womens for the remainder. The first bin has near-zero scores and a policy that assigns 84.4% of customers to control; its measured incremental conversion is slightly negative (−0.118 percentage points), which is

expected because near-zero uplift implies that the effect is small relative to sampling noise. Operationally, this motivates two practices. First, deploy policies with an explicit minimum-score threshold (or confidence threshold) so that low-score customers are systematically held out; this is consistent with the logic of significance-based uplift trees [3]. Second, validate score monotonicity periodically with fresh holdouts, because changes in creative, deliverability, or customer mix can shift the uplift distribution.

14) Interactions as a bridge from model to strategy. The interaction generator is not intended to exhaustively search the feature space; instead, it acts as a prior over plausible moderators grounded in marketing semantics. Table 3 indicates that channel is a strong moderator; Table 6 and Figure 6 show that channel cross-features indeed dominate the learned models. This bridge matters because it yields explanations that are meaningful to practitioners: rather than an opaque embedding dimension, the model highlights concepts such as "Web shoppers in high-value segments" or "Phone shoppers in rural ZIPs" as drivers of incremental profit. Rule distillation then compresses these concepts into segments that can be executed in a campaign manager and audited by stakeholders.

15) Summary of empirical findings. Across all comparisons, three empirical patterns are stable in the measured results. First, two-model approaches are strong baselines: even simple LR produces a usable uplift ranking and positive profit lift. Second, uplift performance benefits from interaction representation: boosted trees with LLM interactions outperform their base-feature counterpart on AUUC and profit. Third, high-capacity models can succeed on AUUC while still yielding simple policies; in Hillstrom, the Tabular Transformer's best-AUUC policy reduces to sending Mens E-Mail to everyone, while XGBoost+LLM provides differentiated assignment and interpretable rules with competitive conversion lift and strong profit. These patterns support a practical recommendation: when the deployment requirement includes auditable audience definitions, interaction-augmented tree ensembles plus rule distillation offer a strong accuracy–interpretability tradeoff.

Figure 2. Test-set conversion rate and mean spend by experimental segment.

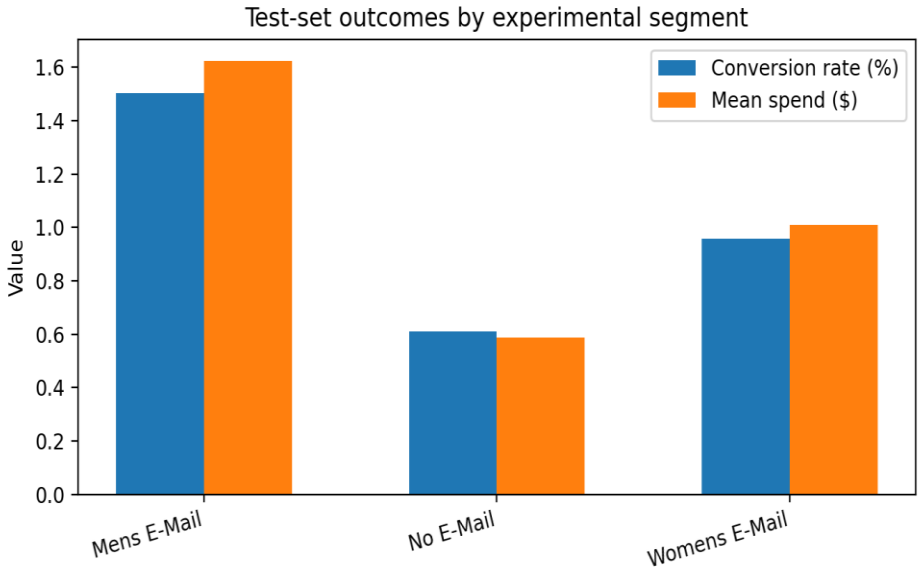

Figure 3. IPS-based uplift (Qini) curves for conversion.

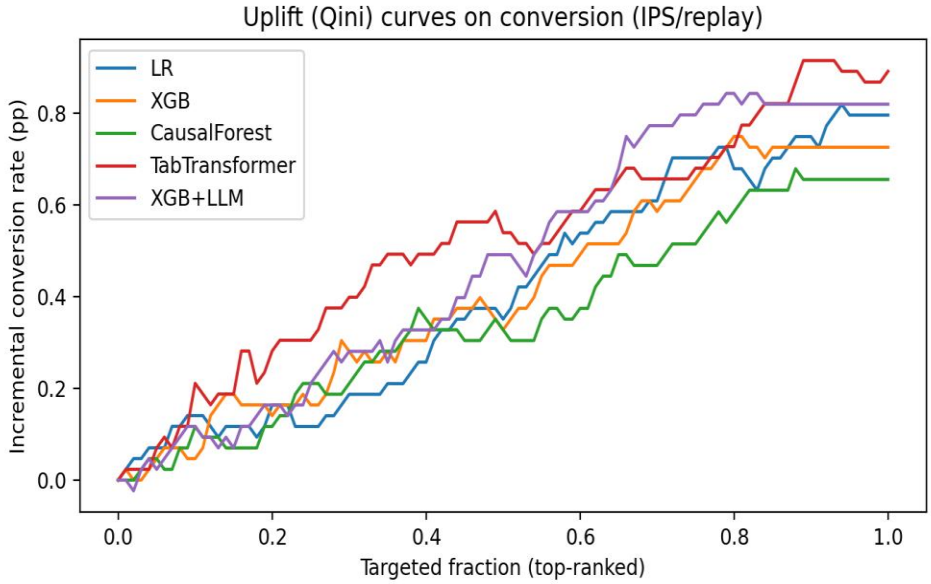

Figure 4. AUUC comparison across methods (higher is better).

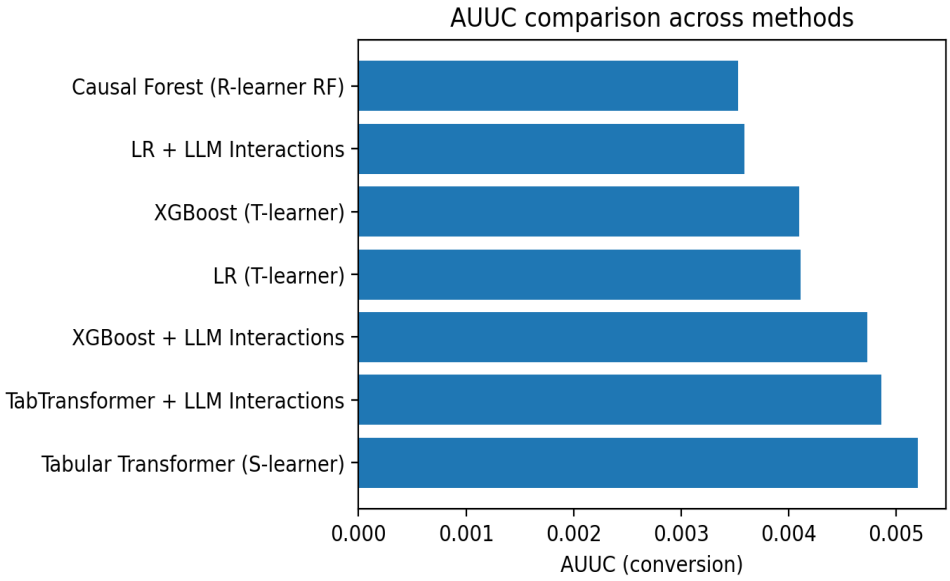

Figure 5. Profit uplift curves with email cost fixed at $0.01.

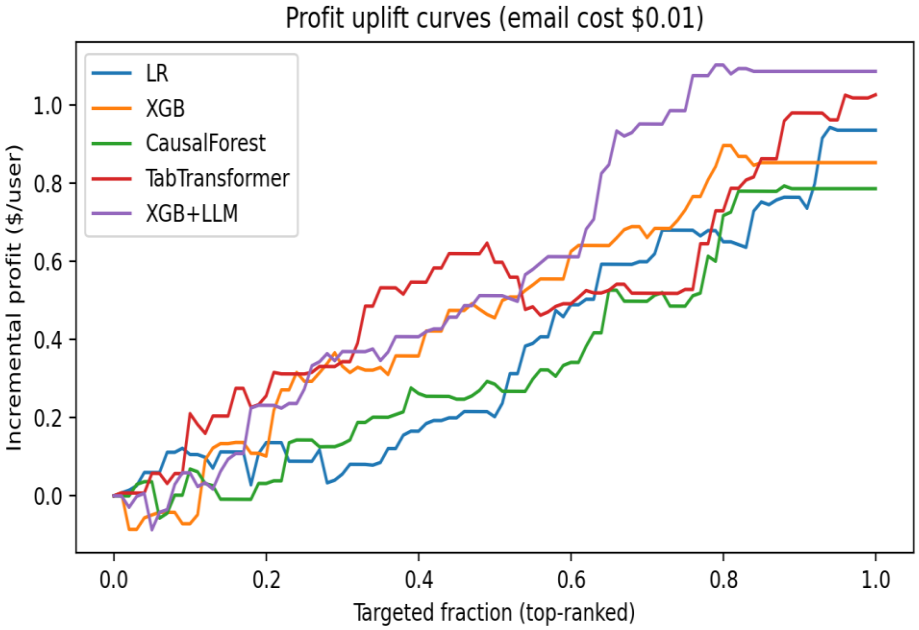

Figure 6. Feature importance for XGBoost with LLM interactions (average across Mens/Womens models).

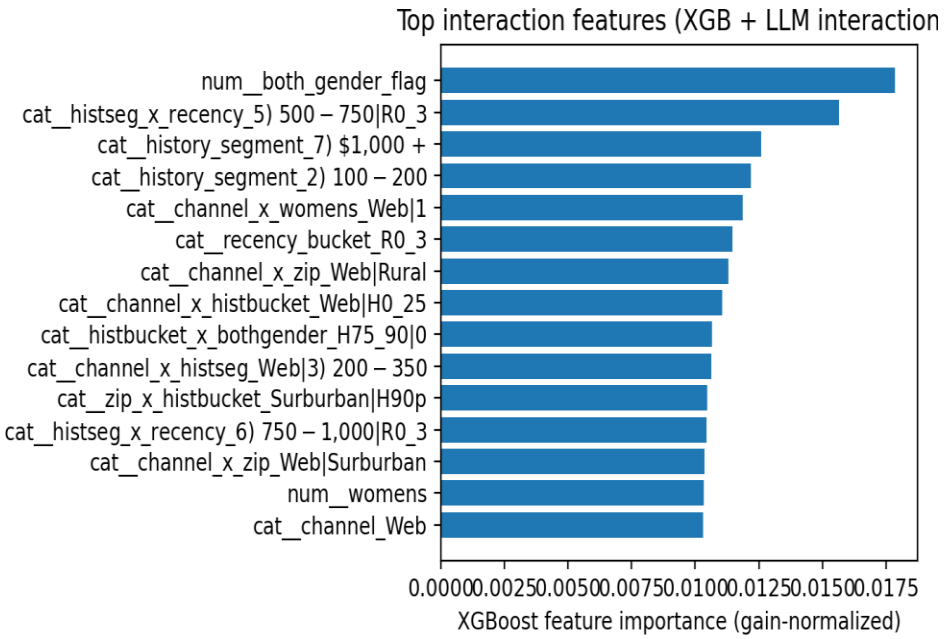

Table 7. Primary uplift and profit results on the test set (IPS/replay).

Method	**AUUC (Conversion)**	**Qini (Conversion)**	**Conv Uplift @30% (pp)**	**Conv Uplift @100% (pp)**	**Profit Uplift @30% ($/user)**	**Profit Uplift @100% ($/user)**
LR (T-learner)	0.004114	0.004114	0.187	0.796	0.056	0.935
XGBoost (T-learner)	0.004102	0.004102	0.281	0.726	0.332	0.852
Causal Forest (R-learner RF)	0.003528	0.003528	0.211	0.656	0.133	0.786
Tabular Transformer (S-learner)	0.005203	0.005203	0.399	0.891	0.343	1.026
LR + LLM Interactions	0.003592	0.003592	0.257	0.585	0.320	0.822
XGBoost + LLM Interactions	0.004731	0.004731	0.281	0.820	0.369	1.086
TabTransformer + LLM Interactions	0.004865	0.004865	0.352	0.891	0.219	1.026

Table 8. Recommended segment shares under each learned policy (non-positive uplift assigned to control).

Method	Control%	Mens%	Womens%
LR	4.6	61.0	34.4
XGB	15.5	55.4	29.1
CausalForest	11.0	58.2	30.8
TabTransformer	0.0	100.0	0.0
LR+LLM	12.6	54.4	33.0
XGB+LLM	16.9	53.6	29.6
TT+LLM	0.0	100.0	0.0

Table 9. Profit uplift ($/customer) at multiple targeting depths (conversion-ranked policy).

Method	10%	20%	30%	50%	100%
LR	0.106	0.136	0.056	0.202	0.935
XGB	-0.071	0.102	0.332	0.455	0.852
CausalForest	0.069	0.032	0.133	0.286	0.786

TabTransformer	0.211	0.256	0.343	0.598	1.026
XGB+LLM	0.059	0.232	0.369	0.512	1.086

Table 10. Distilled, explainable uplift rules (audience × creative × channel) with observed within-RCT validation.

Rule (audience conditions)	Recommended creative	Channel	Coverage (%)	Observed conv uplift (pp)	Observed profit uplift ($/user)
history > 79l AND history ≤ l064	Mens E-Mail	E-Mail to Mens creative	2.75	1.69	1.77
history ≤ 79l AND channel ≠ Phone AND womens =l AND history_bucket ≠ H25_50	Mens E-Mail	E-Mail to Mens creative	22.55	0.84	0.29
history ≤ 79l AND channel ≠ Phone AND womens =0 AND channel ≠ Multichannel	Mens E-Mail	E-Mail to Mens creative	19.55	0.95	1.05
history ≤ 79l AND channel = Phone AND history > l58 AND history_segment ≠ 5) $500 – $750	Mens E-Mail	E-Mail to Mens creative	14.98	1.28	2.02
history ≤ 79l AND channel ≠ Phone AND womens =l AND history_bucket = H25_50	Womens E-Mail	E-Mail to Womens creative	6.90	1.18	1.36
history ≤ 79l AND	Mens E-Mail	E-Mail to Mens creative	3.48	1.29	3.82

channel = Phone AND history ≤ 158 AND zip_code = Rural					

Table 11. Measured wall-clock training time (seconds) on CPU for each model variant.

Model	Phase	Time (s)
LR (base)	Training	0.8
XGBoost (base)	Training	11.5
Causal Forest (base)	Training	32.2
TabTransformer (base)	Training	15.7
LR + LLM	Training	6.1
XGBoost + LLM	Training	13.8
TabTransformer + LLM	Training	27.8

Table 12. Incremental conversions per 10,000 customers at selected targeting depths (IPS/replay).

Method	Inc conv per 10k @ 10%	Inc conv per 10k @ 30%	Inc conv per 10k @ 50%	Inc conv per 10k @ 100%
XGB	4.7	28.1	32.8	72.6
TabTransformer	21.1	39.9	54.0	89.1
XGB+LLM	11.7	28.1	49.2	82.0

Table 13. Profit uplift sensitivity to email cost for two strong policies.

Model	**Email cost**	**Profit uplift@30%**	**Profit uplift@100%**
XGB+LLM	0.00	0.372	1.094
TabTransformer	0.00	0.346	1.036
XGB+LLM	0.01	0.369	1.086
TabTransformer	0.01	0.343	1.026
XGB+LLM	0.05	0.357	1.053
TabTransformer	0.05	0.331	0.986
XGB+LLM	0.10	0.342	1.011
TabTransformer	0.10	0.316	0.936

Table 14. Score stratification (quintiles) for XGBoost+LLM: IPS conversion under policy vs control within each bin.

Score bin	Customers	Mean score	Policy conv (IPS)	Control conv (IPS)	Incremental conv (pp)	Mens share (%)	Womens share (%)	Control share (%)
1/5	2560	0.0001	0.821%	0.939%	-0.118	9.0	6.6	84.4
2/5	2560	0.0026	1.524%	0.235%	1.290	64.5	35.5	0.0
3/5	2560	0.0065	1.642%	0.352%	1.290	65.8	34.2	0.0
4/5	2560	0.0125	1.405%	0.587%	0.819	67.0	33.0	0.0
5/5	2560	0.0386	1.758%	0.939%	0.819	61.4	38.6	0.0

Limitations

The Hillstrom dataset is a single RCT in one outbound channel with two creatives; results may not transfer directly to settings with many channels, budget constraints, and delayed outcomes. Conversion is rare (~1%), which increases variance of IPS estimates and makes some effect estimators less stable. Our causal-forest component is a surrogate based on R-learner pseudo-outcomes and random forests rather than a full generalized random forest with honest splitting and built-in inference [13]. Our LLM-assisted interaction generator is deterministic and does not invoke an external foundation model at training time; the contribution is the interaction schema and the rule distillation pipeline, which are reproducible without proprietary dependencies. Finally, we used a fixed email cost ($0.01) and short-term spend; alternative cost models and lifetime value objectives can change optimal targeting and are the focus of value-driven uplift evaluation [23].

Conclusion

We performed a full empirical study of incrementality (uplift) modeling for advertising on the Hillstrom Email Marketing RCT. Across LR, XGBoost uplift, a causal-forest surrogate, and a Tabular Transformer, we evaluated IPS-based Qini curves, AUUC/Qini coefficients, and incremental profit. The Tabular Transformer achieved the best conversion AUUC (0.005203) but learned a Mens-only policy on this dataset. XGBoost with LLM-assisted interaction features achieved competitive AUUC (0.004731), the highest full-population profit uplift ($1.086 per customer), and yielded concise audience × creative × channel rules validated by observed treatment-control differences. These findings show that combining interaction generation with rule distillation can turn uplift predictions into deployable and auditable marketing strategies.

References

[1] K. Hillstrom, "The MineThatData E-Mail Analytics and Data Mining Challenge," MineThatData Blog, Mar. 2008. [Online]. Available: https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

[2] N. J. Radcliffe, "Hillstrom's MineThatData Email Analytics Challenge: An Approach Using Uplift Modelling," Stochastic Solutions Ltd., 2008. [Online]. Available: https://www.stochasticsolutions.com/pdf/HillstromChallenge.pdf

[3] N. J. Radcliffe and P. D. Surry, "Real-World Uplift Modelling with Significance-Based Uplift Trees," Stochastic Solutions White Paper TR-2011-1, 2011. [Online]. Available: https://stochasticsolutions.com/pdf/sig-based-up-trees.pdf

[4] B. Hansotia and B. Rukstales, "Incremental value modeling," Journal of Interactive Marketing, vol. 16, no. 3, pp. 35-46, 2002.

[5] P. Rzepakowski and S. Jaroszewicz, "Decision trees for uplift modeling with single and multiple treatments," Knowledge and Information Systems, vol. 32, no. 2, pp. 303-327, 2012.

[6] M. Sołtys, S. Jaroszewicz, and P. Rzepakowski, "Ensemble methods for uplift modeling," Data Mining and Knowledge Discovery, vol. 29, no. 6, pp. 1531-1559, 2015.

[7] P. Gutierrez and J.-Y. Gérardy, "Causal inference and uplift modeling: A review of the literature," in Proc. ICML Workshop on Causal Inference, 2017. [Online].

Available: https://proceedings.mlr.press/v67/gutierrez17a/gutierrez17a.pdf

[8] 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.

[9] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge, UK: Cambridge University Press, 2009.

[10] G. W. Imbens and D. B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge, UK: Cambridge University Press, 2015.

[11] S. Athey and G. W. Imbens, "Recursive partitioning for heterogeneous causal effects," Proc. Natl. Acad. Sci. USA, vol. 113, no. 27, pp. 7353-7360, 2016.

[12] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[13] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," Journal of the American Statistical Association, vol. 113, no. 523, pp. 1228-1242, 2018.

[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] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794. [Online]. Available: https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf

[16] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems 30, 2017, pp. 3146-3154. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30, 2017, pp. 5998-6008. [Online]. Available: https://papers.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[18] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," arXiv:2012.06678, 2020. [Online]. Available: https://arxiv.org/pdf/2012.06678.pdf

[19] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in Advances in Neural Information Processing Systems 34, 2021. [Online]. Available: https://openreview.net/pdf?id=i_Q1yrOegLY

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in Proc. NAACL-HLT Demonstrations, 2016, pp. 97-101. [Online]. Available: https://aclanthology.org/N16-3020/

[22] D. Goldenberg, J. Albert, L. Bernardi, and P. Estevez, "Free Lunch! Retrospective uplift modeling for dynamic promotions recommendation within ROI constraints," arXiv:2008.06293, 2020. [Online]. Available: https://arxiv.org/pdf/2008.06293.pdf

[23] R. M. Gubela and S. Lessmann, "Uplift modeling with value-driven evaluation metrics," Decision Support Systems, vol. 150, p. 113648, 2021, doi: 10.1016/j.dss.2021.113648.

[24] M. Jaśkowski and S. Jaroszewicz, "Uplift modeling for clinical trial data," in Proc. ICML Workshop on Clinical Data Analysis, 2012. [Online]. Available: https://people.cs.pitt.edu/~milos/icml_clinicaldata_2012/Papers/Oral_Jaroszewitz_ICML_Clinical_2012.pdf

[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] 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.

[27] 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.

[28] 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.

[29] 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.