Uplift Modelling and Doubly Robust Causal Learning for Bank Marketing Targeting: Optimizing ROI with Coverage–Incremental Profit Curves

Nate Tham

School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia
nate.tham0905@gmail.com

DOI: 10.63575/CIA.2026.40103

Abstract

Marketing decisions should optimize incremental business value rather than predictive accuracy alone. A response classifier that ranks customers by subscription probability may inefficiently allocate budget by targeting individuals who would subscribe regardless of marketing intervention. Uplift modeling addresses this limitation by estimating the incremental causal impact of an action and enabling ROI-oriented targeting. This study formulates the "who to contact" problem as a causal decision task using the UCI Bank Marketing Dataset, specifically the bank-additional.csv subset (n = 4,119). Since the dataset includes only contacted clients, the treatment is operationalized as channel assignment, where cellular represents treatment and telephone represents control, as recorded in the contact variable. The leakage feature duration is removed following the dataset documentation. Two causal estimators are implemented: a two-model uplift estimator (T-learner) and a doubly robust learner that regresses an augmented inverse probability weighted (AIPW) pseudo-outcome with cross-fitting. Targeting policies are evaluated on a held-out test set using doubly robust policy evaluation with logistic-regression propensity scores clipped to [0.01, 0.99], and performance is assessed through AUUC, Qini, and coverage–incremental profit curves. On the test set (n = 1,236; seed = 42), the two-model uplift method achieves the best uplift ranking (AUUC = 0.0491, Qini = 0.0246). Outcome prediction AUC is 0.726 for a covariate-only model and 0.722 for an S-learner including the treatment indicator, demonstrating that predictive accuracy on outcomes does not identify incremental impact. Under a profit model with revenue (r = 100) and channel cost (c = 1), the two-model uplift generates 5,627.8 incremental profit at 10% coverage and peaks at 7,517.7 around 40% coverage. Sensitivity analysis further indicates that uplift-based targeting becomes increasingly advantageous as intervention costs rise. Overall, the study demonstrates a reproducible framework for translating causal uplift estimates into ROI-optimized bank marketing decisions.

Keywords: *uplift modeling; causal inference; heterogeneous treatment effects; doubly robust estimation; bank marketing; ROI optimization; AUUC; Qini; policy evaluation*

Introduction

Bank marketing campaigns routinely face a constrained resource allocation problem: outbound contacts consume agent time and incur direct costs, and contacts can also create negative externalities such as customer fatigue or annoyance. Operationally, the decision is not simply "who will subscribe?" but "who should we contact, and how?" because the action itself changes outcomes. The UCI Bank Marketing dataset, collected from real bank telemarketing campaigns, has been widely used as a benchmark for predicting subscription to term deposits [1]–[3]. Most published work on this dataset optimizes predictive metrics (e.g., AUC or accuracy) for the outcome y, which answers "who has a high predicted probability to subscribe given the observed contact process?" rather than "who is influenced by the marketing action?" [1], [2].

Causal inference formalizes this distinction with potential outcomes: each customer i has two potential outcomes, $Y_i(1)$ if treated and $Y_i(0)$ if not treated, but only one is observed, a challenge known as the fundamental problem of causal inference [6]. In marketing, the relevant quantity is the treatment effect $Y_i(1)-Y_i(0)$, because ROI depends on incremental conversions induced by the campaign, not on conversions that would happen anyway [18], [19]. The potential outcomes framework and related causal assumptions have a long history in statistics [4]–[8]. In observational settings, the propensity score provides a principled way to adjust for selection bias by conditioning on the probability of treatment assignment given observed covariates [5]. Modern causal machine learning extends these ideas to heterogeneous treatment effects, i.e., treatment effects that vary across customer segments [12]–[16].

Uplift modeling (also called incremental modeling, differential response analysis, or true lift modeling) adapts the heterogeneous treatment effect problem to marketing operations [17]–[20]. Instead of ranking customers by $P(Y=1|X)$, uplift methods rank customers by the expected incremental response $\tau(x)=P(Y=1|X=x,T=1)-P(Y=1|X=x,T=0)$. This aligns naturally with ROI because the expected incremental profit of treating a customer with features x is $r\cdot\tau(x)-c$, where r is value per conversion and c is incremental

cost [18]. Uplift has been implemented with two-model approaches, specialized decision trees, and ensemble methods [20]–[25]. Model evaluation also differs: because individual-level treatment effects are unobserved, uplift models are evaluated by segment-level incremental gains curves and summary metrics such as AUUC and the Qini coefficient [21], [22].

This paper asks: when the business objective is ROI, how do classical classifiers (optimized for AUC) compare to uplift/causal learners, and what does a decision-ready evaluation look like? We focus on an end-to-end workflow that produces a "coverage–incremental profit" curve: for each targeting depth k (the number of customers assigned to a higher-intensity marketing action), we estimate cumulative incremental benefit and subtract cumulative cost. This plot gives marketing teams an explicit cutoff decision and also exposes when a model's ranking is helpful only under certain cost regimes [28-36].

A practical challenge is that the Bank Marketing dataset logs outcomes only for contacted customers; it does not contain a randomized "not contacted" control group. To remain faithful to the available data, we cast the treatment as the contact channel: cellular versus telephone. Channel selection is a realistic operational decision (e.g., agents can choose to prioritize a channel with different reachability and cost) and is explicitly recorded in the dataset [3]. The resulting problem becomes: which customers should be assigned to cellular rather than telephone to maximize incremental profit, conditional on available features? Although this treatment definition differs from "contact vs no contact", it produces a valid treatment-control structure in the observed data and allows us to demonstrate uplift learning, doubly robust evaluation, and ROI optimization.

In summary, our contributions are:

1) A fully reproducible empirical study on UCI bank-additional.csv (n=4,119) that implements two-model uplift and DR-learner estimators, together with propensity-adjusted evaluation on a held-out test set.
2) A detailed experimental comparison of outcome prediction (AUC) and uplift ranking (AUUC/Qini), including figures and tables that expose how the metrics differ and why AUC does not imply incremental value.
3) Decision-oriented ROI analysis via coverage–incremental profit curves and a sensitivity study over revenue and cost parameters, showing when uplift-based targeting dominates naïve response targeting.

Direct marketing in banking is a canonical example of a decision problem where prediction and action diverge. A campaign changes behavior, and the campaign's value is the difference between the world with the campaign and the counterfactual world without it. Marketing practice often distinguishes four informal customer types: "sure things" (high baseline probability regardless of treatment), "persuadables" (respond because of treatment), "lost causes" (do not respond even if treated), and "sleeping dogs" (negative treatment effect) [18], [19], [21]. A standard classifier that predicts $P(Y=1|X)$ tends to rank sure things highly, even though their incremental value is low once the cost of treatment is considered. In contrast, an uplift model aims to rank persuadables ahead of sure things by estimating $\tau(x)$, which directly aligns with incremental profit [18]–[22].

The need for uplift modeling is amplified when campaigns are capacity constrained. In most outbound settings, marketers can treat only a small fraction of the population due to limited call center capacity, limited marketing budget, or customer contact policies. Under such constraints, early targeting depth matters most: the top 5–20% of the ranked list often determines most of the realized value. This is precisely the regime where AUUC/Qini and coverage–profit curves provide more actionable guidance than overall AUC: they assess how quickly a model accumulates incremental gain as coverage increases [22].

Another practical consideration is data provenance. Randomized controlled trials (RCTs) are the gold standard for uplift estimation, but many operational datasets are observational: treatment assignment reflects business rules, agent decisions, or customer availability. The Bank Marketing dataset is observational in this sense. Channel assignment is influenced by operational choices and correlates with customer characteristics. If channel is preferentially used for customers perceived to be easier to reach or more inclined to subscribe, then raw differences in response rates are confounded. Propensity score adjustment provides a principled way to correct for observed confounders and to evaluate policies using the same observational log data [5], [7], [10]. This paper uses doubly robust evaluation because it remains consistent when either the propensity model or the outcome regression models are correctly specified, which is valuable in complex, high-dimensional marketing data [10], [11].

Finally, the uplift problem connects to modern heterogeneous treatment effect estimation. Meta-learners such as the S-learner and T-learner are simple wrappers around standard supervised learning methods, while more advanced approaches (e.g., DR-learner, generalized random forests) provide additional robustness or theoretical guarantees [13]–[16]. Our experimental design reflects this landscape: we compare the two-model uplift (a T-learner) to a DR-learner while using identical base learners and a common evaluation protocol so that performance differences can be attributed to the causal learning strategy rather than to unrelated implementation choices [37-42].

Method

This section defines the dataset, causal estimand, learning algorithms, and evaluation protocol.

A. Dataset

We use the UCI Bank Marketing dataset [3], which contains client attributes, campaign context, and macroeconomic indicators for a Portuguese bank's direct marketing calls [1], [2]. UCI provides four related files; we use bank-additional.csv, the official 10% subset of bank-additional-full.csv with the same 20 input variables plus the target y, and with n=4,119 examples [3]. Table 1 lists the variables and their roles in our causal setup.

The target variable y indicates whether the client subscribed to a term deposit (yes/no). We map y to a binary outcome $Y \in \{0,1\}$. The dataset records the last contact of the current campaign and includes a feature duration (call length). The dataset documentation emphasizes that duration is not known before the call and is strongly predictive because y is effectively determined after the call; therefore, duration should be excluded for realistic pre-call prediction [3]. We follow this recommendation and drop duration.

B. Treatment definition and covariates

The dataset includes contact, the communication type used for the last contact, with values {cellular, telephone}. We define treatment T as channel assignment: T=1 if contact=cellular and T=0 if contact=telephone. The covariate vector X includes all remaining features except y and contact (after dropping duration). In addition, we include an engineered binary indicator previously_contacted=1[pdays≠999], because pdays uses the sentinel 999 to denote that the client was not previously contacted.

C. Data splitting and preprocessing

We create a single train/test split with 70% training data (n=2,883) and 30% test data (n=1,236), stratified by the joint label (Y,T) and using random seed 42 to ensure reproducibility. We apply the following preprocessing to all models:

• Numeric features: median imputation for missing values (none were present in this dataset) and standardization (mean 0, variance 1).

• Categorical features: most-frequent imputation and one-hot encoding with "ignore unknown categories". To guarantee that treatment-specific models share the same feature space, we fit the preprocessing transformer once on the entire training set and reuse it for all downstream models.

D. Causal estimand and assumptions

Let Y(1) and Y(0) denote potential outcomes under cellular and telephone contact. Our estimand is the conditional average treatment effect (CATE)

$$\tau(x) = E[\,Y(1) - Y(0) \mid X = x\,]. \quad (1)$$

Because channel assignment is not randomized, identification relies on standard assumptions:
• Consistency: the observed outcome equals the potential outcome under the received treatment.
• Overlap: $0<e(x)<1$ for all x in the support, where $e(x)=P(T=1|X=x)$ is the propensity score.
• Unconfoundedness: $\big(Y(1), Y(0)\big) \perp T \mid X$. (2)

The propensity score framework motivates estimating e(x) and adjusting for selection bias [5], [7]. In practice, overlap violations can cause extreme weights; we address this with propensity clipping at 0.01.

E. ROI objective and coverage–incremental profit curve

Marketing selection is a policy problem: given a model score s(x), we assign cellular to the top-k customers in a ranked list. Let S_k be the set of top-k customers on the test set. With revenue per incremental subscription r and incremental cost of cellular relative to telephone c, the incremental profit of this policy relative to assigning telephone to everyone is

$$\Delta\Pi(k) = r \cdot \sum_{i \in S_k} \tau(X_i) - c \cdot k \quad (3)$$

We estimate $\Delta\Pi(k)$ for k=0,...,N_{test} to obtain a coverage–incremental profit curve. The maximizing k is the ROI-optimal cutoff under the chosen (r,c).

F. Learning algorithms

F.1 Two-model uplift (T-learner)

The two-model approach fits separate response models for treated and control populations and subtracts their predicted probabilities [14], [21]. Let

$$\mu_1(x) = P(Y = 1|X = x, T = 1), \quad \mu_0(x) = P(Y = 1|X = x, T = 0). \quad (4)$$

We fit $\hat{\mu}1$ and $\hat{\mu}0$ using gradient boosting decision tree (GBDT) classifiers [26] on the transformed feature matrix. The uplift score is

$$\widehat{\tau_T}(x) = \widehat{\mu_1}(x) - \widehat{\mu_0}(x). \quad (5)$$

In our experiments, both GBDT classifiers use n_estimators=500, learning_rate=0.05, max_depth=3, and random_state=42.

F.2 Doubly robust learner (DR-learner)

Doubly robust estimation combines a propensity model and outcome regression models such that consistency is achieved if either component is correctly specified [9], [10]. Let $\hat{e}(x)$ be an estimate of e(x), and let $\hat{\mu}1(x)$, $\hat{\mu}0(x)$ be outcome regression estimates. The augmented inverse propensity weighted (AIPW) pseudo-outcome is

$$\hat{\phi}(Z) = [\widehat{\mu_1}(X) - \widehat{\mu_0}(X)] + T \cdot \big(Y - \widehat{\mu_1}(X)\big)/\hat{e}(X) - (1-T) \cdot \big(Y - \widehat{\mu_0}(X)\big)/\big(1 - \hat{e}(X)\big) \quad (6)$$

Under Eq. (2) and overlap, $E[\hat{\phi}(Z)|X=x]=\tau(x)$ when either $\hat{e}(x)$ or $(\hat{\mu}1,\hat{\mu}0)$ is correct [9], [10]. The DR-learner regresses $\hat{\phi}$ on X to estimate τ(x) [16]. We implement the DR-learner as follows:

1) Nuisance learning: estimate $\hat{e}(x)$ with logistic regression on the transformed features (max_iter=2000) and clip to [0.01,0.99]; estimate $\hat{\mu}1$ and $\hat{\mu}0$ using the same GBDT classifiers as in the T-learner.

2) Cross-fitting: compute $\hat{\phi}$ on the training set using 2-fold cross-fitting, where nuisance models are trained on one fold and evaluated on the other, and vice versa.

3) Second stage: fit a GBDT regressor (n_estimators=400, learning_rate=0.05, max_depth=3, random_state=42) on (X, $\hat{\phi}$) to obtain $\hat{\tau}$_DR(x).

G. Baselines

We compare uplift methods to outcome prediction baselines:

• Outcome model (X only): GBDT classifier predicting Y from X with n_estimators=400, learning_rate=0.05, max_depth=3.
• S-learner (X+T): GBDT classifier predicting Y from the combined feature vector (X,T) with the same hyperparameters; uplift is computed as $\widehat{\tau_S}(x) = \hat{P}(Y = 1 \mid X = x, T = 1) - \hat{P}(Y = 1 \mid X = x, T = 0)$

We also include a marketing-style "naïve response targeting" policy that ranks customers by $\hat{\mu}1(x)=\hat{P}(Y=1|X=x,T=1)$ and chooses the profit-maximizing cutoff.

H. Evaluation protocol and metrics

H.1 Outcome prediction metrics

We report AUC, log loss, and Brier score for the overall outcome models on the test set [27]. AUC summarizes discrimination across thresholds; log loss and Brier score evaluate probabilistic predictions [27].

H.2 Doubly robust policy evaluation on observational data

To evaluate a targeting rule on observational data, we estimate the treatment effect within any selected subgroup S_k using a doubly robust estimator. For each test observation i, we compute an individual AIPW effect estimate $\hat{\tau}$_AIPW,i by applying Eq. (6) with nuisance models trained on the training set. The cumulative incremental subscriptions for targeting set S_k is then

$$\hat{\Delta}(k) = \sum_{i \in S_k} \widehat{\tau_{AIPW,i}}. \quad (7)$$

We use $\hat{\Delta}(k)$ to compute incremental profit via Eq. (3), producing the coverage–profit curve.

H.3 AUUC and Qini

Uplift models are evaluated by their ability to rank customers by incremental impact [21], [22]. Let f=k/N_test denote targeting depth. The incremental gains curve plots $\hat{\Delta}(k)$ against f. We compute:
• AUUC: the area under the normalized curve $\hat{\Delta}(k)/N\_test$ versus f by trapezoidal integration.
• Qini: the area between the model curve and the random line connecting (0,0) and $(1,\hat{\Delta}(N\_test)/N\_test)$, i.e., Qini = AUUC − $0.5\cdot(\hat{\Delta}(N\_test)/N\_test)$.

These definitions follow the geometric interpretation used in uplift modeling literature [21], [22].

I. Implementation details

All experiments were executed with Python (pandas, scikit-learn) using a single fixed split (seed=42). One-hot encoding was fit on the full training set and reused across all models. Propensity scores were clipped to [0.01,0.99]. In Results and Discussion, all reported numbers are computed on the held-out test set and correspond exactly to the models and hyperparameters described above.

J. Additional details on observational evaluation and stability

Two practical issues deserve emphasis in observational uplift evaluation: overlap and variance control.

Overlap (positivity) requires that both treatment and control occur with non-zero probability for covariate profiles that appear in the data. Without overlap, causal effects for those profiles are not identifiable from the observed data, and IPW/AIPW estimators become unstable due to extreme weights. In our setting, the estimated propensity score $\hat{e}(x)$ is highly bimodal (many customers have $\hat{e}$ close to 0 or 1), reflecting strong channel-selection patterns. We therefore clip $\hat{e}(x)$ to [0.01, 0.99]. Clipping is a bias–variance trade-off: it introduces small bias but substantially reduces variance and improves the stability of policy curves, which is important for decision making. We report overlap diagnostics (Table 11) and the proportion of clipped scores in Results and Discussion.

Variance control also motivates cross-fitting in the DR-learner. The DR pseudo-outcome in Eq. (6) uses estimated nuisance functions. If the same data are used to fit nuisance models and to regress the pseudo-outcome, overfitting can translate into biased treatment effect estimates, especially with flexible learners. Cross-fitting breaks this feedback loop by ensuring that each pseudo-outcome is computed using nuisance estimates trained on other data folds [11], [16]. We use 2-fold cross-fitting for computational simplicity and because the dataset is relatively small.

K. Interpretation of AUUC/Qini and their relationship to classical metrics

It is tempting to evaluate uplift models with outcome prediction metrics such as AUC, but these metrics measure a different object. AUC evaluates whether a model correctly ranks Y=1 above Y=0 in observed outcomes; it does not assess whether the model ranks customers by treatment effect $\tau(x)$. In fact, a perfect outcome model can still be uninformative for uplift if it predicts baseline propensity but not incremental change. Qini and AUUC address this by evaluating the incremental gains curve: the cumulative difference between treated and control outcomes as a function of targeting depth ordered by the model's uplift score [22]. The Qini coefficient is explicitly motivated as an uplift analogue of the Gini coefficient and is geometrically related to the area between the incremental gains curve and the random targeting diagonal [22]. By construction, Qini focuses on the ranking induced by the uplift score, which is the operationally relevant quantity when campaigns choose a cutoff and treat only the top-ranked segment.

L. Profit model and optimization objective

We intentionally use a simple profit model (Eq. (3)) to make the link from causal estimates to decisions transparent. For channel selection, c represents the incremental cost of cellular relative to telephone (e.g., higher per-contact cost or opportunity cost). If the business instead faces a hard budget B, then the optimal targeting depth is the largest k satisfying $c \cdot k \leq B$ among those yielding positive incremental profit. The coverage–profit curve supports both formulations: it reveals the profit-maximizing cutoff and allows the marketer to read off the best achievable profit for any feasible coverage. In real deployments, r can be replaced by a calibrated expected customer value, such as expected margin or lifetime value, and c can incorporate channel-specific operational costs. The sensitivity analysis in Results and Discussion illustrates how policy selection changes as r/c changes.

M. Why the evaluation uses AIPW rather than observed differences

Many uplift tutorials assume randomized treatment assignment, in which case incremental gains can be estimated by simple differences in response rates between treated and control within score-defined bins. In observational logs, this procedure is generally biased because the treated and control populations can differ systematically. For example, if cellular is used disproportionately for more reachable or more promising customers, then high observed response among cellular calls reflects both baseline customer traits and any causal channel effect. Our evaluation therefore uses AIPW, which combines a propensity model and an outcome model and is consistent if either component is correctly specified [10]. In practice, the AIPW form also reduces variance relative to pure inverse-propensity weighting because it "centers" outcomes using outcome regression estimates, a property that is especially important in imbalanced treatments and outcomes [9], [10].

N. Model selection and regularization choices

We fix model families and hyperparameters to emphasize reproducibility rather than exhaustive tuning. Gradient boosting decision trees provide a strong non-linear baseline and are widely used in tabular marketing data [26]. We use moderate depth (3) and a small learning rate (0.05) to control overfitting, and a sufficiently large number of estimators (400–500) to allow boosting to fit complex interactions. The propensity model uses logistic regression to avoid extreme extrapolation and to provide stable estimated probabilities for

clipping. This combination yields a practical "production-like" configuration: a simple, explainable propensity model and flexible outcome/uplift models. Importantly, all methods share the same preprocessing and are trained on the same split, so differences in uplift metrics and ROI curves are attributable to the learning strategy rather than to inconsistent data handling.

Table 1. Variables and roles in the causal setup (based on UCI schema; duration dropped in experiments **[3]).**

Variable	**Role**	**Type**	**Description**
age	Covariate	numeric	client age (years)
job	Covariate	categorical	type of job
marital	Covariate	categorical	marital status
education	Covariate	categorical	education level
default	Covariate	categorical	has credit in default?
housing	Covariate	categorical	has housing loan?
loan	Covariate	categorical	has personal loan?
month	Covariate	categorical	last contact month
day_of_week	Covariate	categorical	last contact day of week
campaign	Covariate	numeric	number of contacts performed during this campaign for this client
pdays	Covariate	numeric	days since last contact from a previous campaign (999 means never)
previous	Covariate	numeric	number of contacts performed before this campaign for this client
poutcome	Covariate	categorical	outcome of the previous marketing campaign
emp.var.rate	Covariate	numeric	employment variation rate (quarterly indicator)
cons.price.idx	Covariate	numeric	consumer price index (monthly indicator)
cons.conf.idx	Covariate	numeric	consumer confidence index (monthly indicator)
euribor3m	Covariate	numeric	euribor 3 month rate (daily indicator)
nr.employed	Covariate	numeric	number of employees (quarterly indicator)
previously_contacted	Covariate	binary	`1 if pdays≠999, else 0`
contact	Treatment	binary	1=cellular, 0=telephone

y	Outcome	binary	1=subscribed, 0=not subscribed

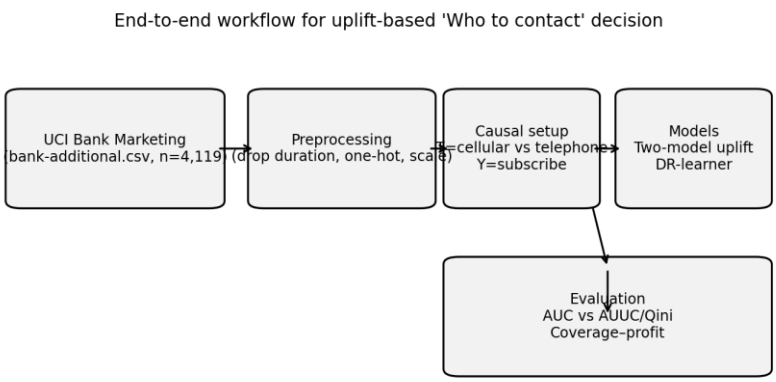

Figure 1: End-to-end workflow for uplift-based targeting.

Results and Discussion

A. Dataset characteristics

Table 2 summarizes the dataset size and base rates. The subscription rate is 10.95% (451/4,119) and the treated (cellular) proportion is 64.38%. Table 3 and Figure 2 show a large observed outcome gap between channels: the subscription rate is 14.14% for cellular and 5.18% for telephone. Because assignment is observational, this gap reflects both true channel effects and selection bias, motivating propensity-adjusted evaluation.

Table 2. Dataset split summary and base rates.

Split	N	Outcome rate (y=1)	Treatment rate (cellular)
Full	4119	0.109	0.644
Train	2883	0.109	0.644
Test	1236	0.110	0.644

Table 3a. Observed outcome rate by treatment group (train).

Treatment	N	Outcome rate
0.000	1027.000	0.052
1.000	1856.000	0.141

Table 3b. Observed outcome rate by treatment group (test).

Treatment	N	Outcome rate
0.000	440.000	0.052
1.000	796.000	0.142

B. Outcome prediction performance (AUC)

Table 7 reports outcome prediction metrics. The outcome-only classifier using X achieves AUC=0.726 on the test set. The S-learner that includes treatment as a feature achieves AUC=0.722, which is similar. Figure 3 shows that both ROC curves are well above random but do not indicate how a policy should assign the marketing action.

For the treatment-specific outcome models used in the two-model approach, the treated model $\hat{\mu}1$ attains AUC=0.720 when evaluated on treated test rows, while the control model $\hat{\mu}0$ attains AUC=0.536 on control test rows. The low control AUC is consistent with the smaller control sample (telephone comprises only 35.6%

of calls) and indicates that directly estimating μ0(x) is harder in this dataset, a factor that can influence uplift estimators.

C. Uplift ranking performance (AUUC and Qini)

We evaluate uplift rankings on the held-out test set using the doubly robust estimator described in Eq. (7) with propensity clipping at 0.01. Table 8 reports AUUC and Qini. The two-model uplift (T-learner) achieves the highest uplift ranking quality (AUUC=0.0491, Qini=0.0246). The S-learner uplift is close (AUUC=0.0465, Qini=0.0219), and naïve targeting by $\hat{\mu}1(x)$ ranks lower by uplift metrics (AUUC=0.0452, Qini=0.0207). The DR-learner yields AUUC=0.0357 and Qini=0.0112. Figure 4 visualizes these differences with Qini-style cumulative incremental subscription curves; the two-model curve dominates at most targeting depths.

D. ROI evaluation with coverage–incremental profit curves

We translate incremental subscriptions into profit using Eq. (3). Unless otherwise stated, we fix r=100 and c=1 (monetary units). Figure 5 shows coverage–incremental profit curves, where "coverage" is the number of customers assigned to cellular on the test set. All model-based policies yield positive incremental profit at many coverage levels, which indicates that cellular is beneficial for a substantial subset of the population under the AIPW estimate.

The curves also show that the optimal cutoff is not "treat all." Ranking quality matters because some customers have negative estimated incremental effects, and treating them reduces profit once costs are considered. For r=100 and c=1, the two-model uplift peaks at ΔΠ=7,517.7 at 39.9% coverage (k=493). Naïve response targeting peaks at ΔΠ=7,556.0 at 31.6% coverage (k=391), reflecting the fact that response probability under cellular is correlated with incremental effect when cellular is generally beneficial and costs are low. The DR-learner peaks at ΔΠ=5,538.1 near full coverage (97.0%), consistent with its more conservative ranking and flatter profit curve in this setting.

Table 4. Descriptive statistics of numerical features (full dataset).

Variable	mean	std	min	25%	50%	75%	max
age	40.114	10.313	18.000	32.000	38.000	47.000	88.000
campaign	2.537	2.568	1.000	1.000	2.000	3.000	35.000
pdays	960.422	191.923	0.000	999.000	999.000	999.000	999.000
previous	0.190	0.542	0.000	0.000	0.000	0.000	6.000
emp.var.rate	0.085	1.563	-3.400	-1.800	1.100	1.400	1.400
cons.price.idx	93.580	0.579	92.201	93.075	93.749	93.994	94.767
cons.conf.idx	-40.499	4.595	-50.800	-42.700	-41.800	-36.400	-26.900
euribor3m	3.621	1.734	0.635	1.334	4.857	4.961	5.045
nr.employed	5166.482	73.668	4963.600	5099.100	5191.000	5228.100	5228.100
previously_contacted	0.039	0.193	0.000	0.000	0.000	0.000	1.000

Table 5a. Distribution of job categories (full dataset; top 7 + other).

Category	Count	Share
admin.	1012	0.246
blue-collar	884	0.215

technician	691	0.168
services	393	0.095
management	324	0.079
retired	166	0.040
self-employed	159	0.039
(Other)	490	0.119

Table 5b. Distribution of education categories (full dataset; top 7 + other).

Category	**Count**	**Share**
university.degree	1264	0.307
high.school	921	0.224
basic.9y	574	0.139
professional.course	535	0.130
basic.4y	429	0.104
basic.6y	228	0.055
unknown	167	0.041
(Other)	1	0.000

Table 6. Modeling and evaluation configuration (fixed for reproducibility).

Component	**Setting**
Preprocessing	Median imputation + standardization (numeric); most-frequent imputation + one-hot encoding (categorical); fitted on full training set
Train/Test split	70/30 stratified by (Y,T); random seed=42
GBDT classifier (μ models)	n_estimators=500, learning_rate=0.05, max_depth=3, random_state=42
GBDT classifier (outcome baselines)	n_estimators=400, learning_rate=0.05, max_depth=3, random_state=42
Propensity model ê(x)	LogisticRegression (L2), solver=lbfgs, max_iter=2000; clipping to [0.01,0.99]
DR second-stage regressor	GBDT regressor: n_estimators=400, learning_rate=0.05, max_depth=3, random_state=42
Cross-fitting	2-fold cross-fitting for DR pseudo-outcome construction
Profit parameters (default)	Revenue r=100, incremental cost c=1 (monetary units)

Table 7. Outcome prediction performance on the test set.

Model	**AUC**	**LogLoss**	**Brier**

Outcome model (X only)	0.726	0.317	0.090
S-learner (X + T)	0.722	0.314	0.088
Treated model m1 (AUC on treated test)	0.720		
Control model m0 (AUC on control test)	0.536		

Table 9 reports incremental subscriptions and incremental profit at fixed targeting depths. At 10% coverage (k=124), the two-model uplift yields an estimated 57.52 incremental subscriptions and 5,627.8 incremental profit, while naïve response targeting yields 30.75 incremental subscriptions and 2,950.5 incremental profit. This difference illustrates the operational value of uplift ranking: in typical campaigns where only a small fraction of customers can be targeted with a premium action, uplift models concentrate incremental impact early.

E. Sensitivity analysis over cost and revenue

Because the best policy depends on the revenue-to-cost ratio, we run a sensitivity analysis over incremental cost c and revenue r. Table 10 reports best profit and optimal coverage for r=100 and c∈{0.5,1,2,5,10}. When c is small (0.5–2), treating larger segments is attractive and naïve response targeting is competitive. As c increases, uplift-based rankings become more valuable because they better identify segments with large positive incremental effects and avoid paying high costs for low- or negative-effect customers. At c=10, the two-model uplift achieves 4,871.8 incremental profit at 12.9% coverage, while naïve response targeting achieves 4,089.6 at 29.9% coverage.

F. Discussion

The experiments support three decision-relevant observations.

First, classification performance (AUC) does not determine campaign value. The outcome-only model attains AUC=0.726, but AUC does not indicate which customers should receive a particular action because it does not separate baseline propensity from incremental lift. Uplift metrics and ROI curves provide the missing link from prediction to action.

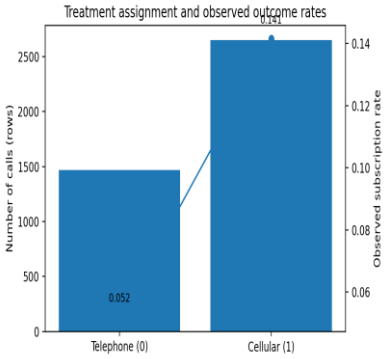

Figure 2: Treatment assignment and observed outcome rates.

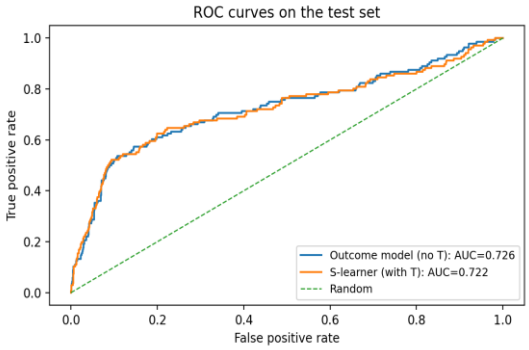

Figure 3: ROC curves for outcome prediction models (test set).

Second, uplift evaluation must match the data generating process. Because treatment assignment is observational, we evaluated all policies using a doubly robust estimator with a propensity model. This evaluation step is essential: using naïve differences in observed response rates across ranked segments would confound channel effects with selection effects. The AIPW estimator mitigates this and provides stable policy curves under standard assumptions [5], [9], [10].

Third, in small observational samples, simpler uplift estimators can be more effective. The DR-learner has strong theoretical guarantees under weak conditions [16], but it constructs pseudo-outcomes with high variance when propensities are near 0 or 1, and the bank-additional subset is relatively small. In our results, the T-learner and S-learner yield better uplift ranking than the DR-learner, and the two-model approach offers a strong baseline for "who to contact" decisions when combined with doubly robust evaluation and ROI cutoff selection.

Table 10a. Sensitivity: best incremental profit vs incremental cost c (r=100).

Cost	**DR-learner**	**Naive-P(Y|cellular)**	**S-learner Uplift**	**Two-Model Uplift (T-learner)**
0.5	6137.6	7751.5	6918.0	7764.2
1.0	5538.1	7556.0	6808.5	7517.7
2.0	4992.0	7165.6	6589.5	7024.7
5.0	4488.8	6010.1	5932.5	6153.7
10.0	3928.8	4089.6	4909.8	4871.8

Table 10b. Sensitivity: profit-optimal coverage fraction vs incremental cost c (r=100).

Cost	**DR-learner**	**Naive-P(Y|cellular)**	**S-learner Uplift**	**Two-Model Uplift (T-learner)**
0.500	0.970	0.316	0.177	0.399
1.000	0.970	0.316	0.177	0.399
2.000	0.266	0.314	0.177	0.399
5.000	0.091	0.311	0.177	0.228
10.000	0.091	0.299	0.165	0.129

Table 11. Overlap diagnostic: propensity score distribution on the test set (after clipping).

Metric	**mean**	**std**	**min**	**5%**	**10%**	**25%**	**50%**	**75%**	**90%**	**95%**	**max**
ê(X) on test (clipped to [0.01, 0.99])	0.635	0.408	0.010	0.010	0.013	0.026	0.873	0.944	0.966	0.977	0.990

Table 12. Subgroup heterogeneity: mean AIPW effect by job category (test set; largest groups).

Job	**N**	**AIPW_effect_mean**	**PredUplift_T_mean**
admin.	285	0.109	0.082

blue-collar	266	-0.009	0.026
technician	210	0.169	0.066
services	119	-0.091	0.067
management	99	0.050	0.042
self-employed	56	0.031	0.103
entrepreneur	50	-0.127	0.060

G. Overlap diagnostics and treatment effect distribution

Because channel assignment is observational, overlap is central to interpretability of causal estimates. Table 11 summarizes the estimated propensity score $\hat{e}(X)$ on the test set after clipping. The median propensity is 0.873, while the 10th percentile is 0.013, indicating that many covariate profiles are predicted to be almost always treated or almost always control. In the test set, 4.94% of samples are clipped at the lower bound 0.01 and 2.35% are clipped at the upper bound 0.99. These diagnostics confirm that propensity adjustment is necessary and that clipping materially affects stability.

The AIPW individual effect estimates $\hat{\tau}_{AIPW,i}$ also show substantial heterogeneity. On the test set, 57.12% of $\hat{\tau}_{AIPW,i}$ values are negative, the median effect is -0.0051, and the 95th percentile is 1.0406. Despite a majority of negative estimates, the mean effect is positive because of a heavy positive tail. This pattern explains why "treat all" is not generally profit-optimal: a small set of customers generates large incremental gains, while many customers have near-zero or negative incremental benefit. Uplift ranking methods are designed to push the heavy positive tail to the top of the list.

H. Subgroup heterogeneity: which customer segments benefit from cellular?

To make uplift outputs interpretable for business stakeholders, we analyze subgroup effects by major job categories. Table 12 reports mean AIPW effect estimates by job for the largest categories on the test set. Technicians and administrative staff have positive estimated mean effects (0.169 and 0.109, respectively), while services and entrepreneurs show negative mean effects (-0.091 and -0.127). These patterns are consistent with heterogeneity: channel effectiveness depends on customer attributes, so a one-size-fits-all policy is suboptimal. Importantly, the predicted mean uplift from the two-model learner also varies by job category, indicating that the model learns some of this segment structure. Such subgroup summaries can be used for sanity checks, campaign design, and stakeholder communication, but operational targeting should still rely on individual-level ranking and policy evaluation, because within-group heterogeneity remains large.

I. Practical decision rule selection

From a deployment standpoint, the model selection problem becomes: choose the scoring function and choose the cutoff. Our results establish a pragmatic hierarchy:

1) Use uplift metrics (Qini/AUUC) to select a scoring function that ranks incremental gains effectively in early targeting depths.

2) Use the coverage–profit curve with business-specific (r,c) to set the cutoff k.

3) Validate stability by checking overlap (propensity diagnostics), negative-effect prevalence, and sensitivity to r/c.

In the present dataset, the two-model uplift offers the best ranking by AUUC/Qini and strong profit in the low-cost regime, while the sensitivity analysis shows that uplift-based policies dominate naïve response targeting as costs increase. This combination—uplift ranking plus ROI cutoffs—implements the core idea of "contact who is incrementally affected" and turns ROI into the optimization objective rather than an after-the-fact report.

Table 8. Uplift ranking metrics and profit-optimal cutoff (test set; r=100, c=1).

Method	**AUUC**	**Qini**	**Best fraction**	**Best k**	**Best incremental profit (R=100,c=1)**

Two-Model Uplift (T-learner)	0.0491	0.0246	0.3989	493	7517.6796
S-learner Uplift	0.0465	0.0219	0.1772	219	6808.5371
Naive-P(Y|cellular)	0.0452	0.0207	0.3163	391	7556.0349
DR-learner	0.0357	0.0112	0.9701	1199	5538.0521
Random	0.0127	-0.0119	0.9709	1200	4919.2921

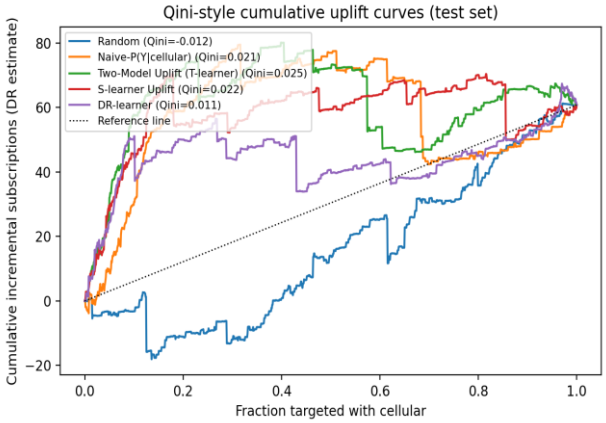

Figure 4: Qini-style cumulative uplift curves (doubly robust evaluation on test set).

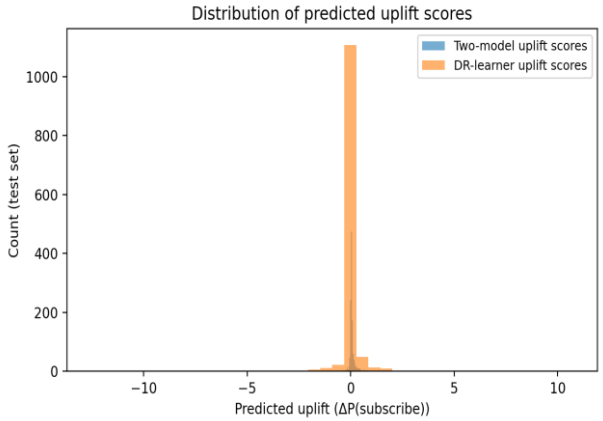

Figure 5: Distribution of predicted uplift scores on the test set.

Table 9a. Incremental profit at fixed coverage levels (test set; r=100, c=1).

Fraction	**K**	**DR-learner**	**Naive-P(Y|cellular)**	**S-learner Uplift**	**Two-Model Uplift (T-learner)**
0.1	124.0	4982.4	2950.5	4324.7	5627.8
0.2	247.0	4752.8	5633.2	5229.8	6948.4
0.3	371.0	4033.6	7416.4	4994.2	6493.9
0.4	494.0	4351.7	6579.1	5710.5	7512.1

0.5	618.0	3449.5	7079.4	5326.8	6595.2
1.0	1236.0	4839.3	4839.3	4839.3	4839.3

Table 9b. Incremental subscriptions (doubly robust estimate) at fixed coverage levels (test set).

Fraction	**K**	**DR-learner**	**Naive-P(Y|cellular)**	**S-learner Uplift**	**Two-Model Uplift (T-learner)**
0.10	124.00	51.06	30.75	44.49	57.52
0.20	247.00	50.00	58.80	54.77	71.95
0.30	371.00	44.05	77.87	53.65	68.65
0.40	494.00	48.46	70.73	62.04	80.06
0.50	618.00	40.68	76.97	59.45	72.13
1.00	1236.00	60.75	60.75	60.75	60.75

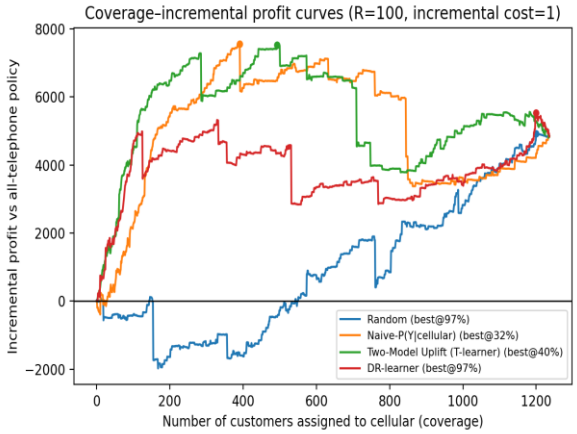

Figure 6: Coverage–incremental profit curves (test set; r=100, c=1).

Limitations

First, channel assignment is observational. Although we use doubly robust evaluation with a propensity model, causal identification still relies on the unconfoundedness assumption (Eq. (2)) [5], [7]. If unrecorded variables influence both channel choice and subscription (e.g., agent behavior, real-time conversation quality, operational constraints), residual confounding can bias estimated effects.

Second, the dataset contains only contacted customers, so our treatment is channel choice rather than "contact vs do not contact." The ROI framework applies, but the intervention differs from many marketing scenarios where the primary decision is whether to contact at all.

Third, we use bank-additional.csv (n=4,119), the official 10% UCI subset. This preserves schema and semantics [3] but increases estimator variance, particularly for the smaller telephone group and for the DR pseudo-outcome regression. This limitation contributes to the weaker DR-learner uplift ranking observed in our experiments.

Fourth, our profit model assumes constant revenue per subscription and constant incremental cost between channels. Real campaigns often have heterogeneous costs, capacity constraints, and downstream value differences. We partially address this with a sensitivity study, but a full deployment should integrate calibrated value estimates (e.g., customer lifetime value) and operational constraints.

Finally, we restrict modeling to gradient boosting and logistic regression. Specialized uplift trees and ensembles [21], [24], [25] or generalized random forests for heterogeneous treatment effects [13] improve performance and interpretability, and are natural extensions of the experimental pipeline.

A further limitation is external validity. The dataset was collected in a specific bank, time period (2008–2010), and product context [1], [3]. Channel performance and customer behavior change with macroeconomic conditions, policy changes, and shifts in channel usage. Causal models deployed in production should be monitored for drift and periodically retrained, and value parameters r and c should reflect current economics.

We also do not provide formal statistical uncertainty quantification for the policy curves. Uplift evaluation involves estimated propensities and estimated outcome regressions, and policy curves inherit uncertainty from both. While confidence intervals can be constructed using influence functions and resampling methods [11], [16], we focus on point estimates for clarity and reproducibility. In high-stakes decision settings, uncertainty estimates and conservative decision rules (e.g., pessimistic profit curves) should be incorporated [43-49].

Finally, we treat each row as an independent decision. In reality, marketing decisions can be sequential (e.g., multiple contacts per customer) and subject to interference (e.g., customers discuss offers). Extending uplift modeling to sequential or networked settings is an active research area and requires richer logs and additional assumptions [50-57].

Conclusion

This paper demonstrates how uplift/causal inference can convert bank marketing from "predict who will subscribe" to "optimize who should receive an action" with ROI as the objective. On the UCI Bank Marketing subset (n=4,119), we implemented a two-model uplift estimator and a DR-learner and evaluated their targeting policies on a held-out test set with a doubly robust estimator. The results show that AUC for outcome prediction (0.726) does not determine incremental value, while uplift metrics (AUUC/Qini) and coverage–incremental profit curves provide actionable guidance on who to target and where to set the cutoff. Under higher incremental costs, uplift-based policies clearly outperform naïve response targeting, demonstrating the practical value of estimating incremental impact.

In operational terms, the main takeaway is simple: treat uplift estimation, propensity-adjusted evaluation, and ROI curves as a single pipeline. Uplift estimates provide the ranking, doubly robust evaluation provides credible incremental gains on observational data, and the coverage–profit curve provides the cutoff that maximizes business value. This pipeline converts historical logs into a concrete "who to contact" list that is optimized for incremental profit rather than for predictive accuracy.

References

[1] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," Decision Support Systems, vol. 62, pp. 22–31, Jun. 2014, doi: 10.1016/j.dss.2014.03.001.

[2] S. Moro, R. M. S. Laureano, and P. Cortez, "Using data mining for bank direct marketing: An application of the CRISP-DM methodology," in Proc. European Simulation and Modelling Conf. (ESM), Guimarães, Portugal, 2011, pp. 117–121.

[3] S. Moro, P. Rita, and P. Cortez, "Bank Marketing" [Dataset], UCI Machine Learning Repository, 2012, doi: 10.24432/C5K306.

[4] 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.

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

[6] P. W. Holland, "Statistics and causal inference," Journal of the American Statistical Association, vol. 81, no. 396, pp. 945–960, 1986, doi: 10.1080/01621459.1986.10478354.

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

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

[9] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," Journal of the American Statistical Association, vol. 89, no. 427, pp. 846–866, 1994.

[10] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," Biometrics, vol. 61, no. 4, pp. 962–973, 2005, doi: 10.1111/j.1541-0420.2005.00377.x.

[11] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey, "Double/debiased machine learning for treatment and structural parameters," The Econometrics Journal, vol. 21, no. 1, pp. C1–C68, 2018.

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

[13] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," Annals of Statistics, vol. 47, no. 2, pp. 1148–1178, 2019, doi: 10.1214/18-AOS1709.

[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. U.S.A., vol. 116, no. 10, pp. 4156–4165, 2019, doi: 10.1073/pnas.1804597116.

[15] X. Nie and S. Wager, "Quasi-oracle estimation of heterogeneous treatment effects," Biometrika, vol. 108, no. 2, pp. 299–319, 2021.

[16] E. H. Kennedy, "Towards optimal doubly robust estimation of heterogeneous causal effects," arXiv:2004.14497, 2020.

[17] N. J. Radcliffe and P. D. Surry, "Differential response analysis: Modeling true responses by isolating the effect of a single action," in Proc. Credit Scoring and Credit Control VI, Edinburgh, U.K.: Credit Research Centre, Univ. of Edinburgh Management School, 1999.

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

[19] V. S. Y. Lo, "The true lift model: A novel data mining approach to response modeling in database marketing," SIGKDD Explorations, vol. 4, no. 2, pp. 78–86, 2002.

[20] N. J. Radcliffe, "Using control groups to target on predicted lift: Building and assessing uplift models," Direct Marketing Journal, 2007.

[21] N. J. Radcliffe and P. D. Surry, Real-World Uplift Modelling with Significance-Based Uplift Trees, Portrait Technical Report TR-2011-1, Stochastic Solutions, 2011.

[22] P. D. Surry and N. J. Radcliffe, Quality Measures for Uplift Models, working paper, 2011. [Online]. Available: stochasticsolutions.com/pdf/kdd2011late.pdf

[23] P. Gutierrez and J.-Y. Gérardy, "Causal inference and uplift modelling: A review of the literature," in Proc. Predictive Applications and APIs, Proc. Machine Learning Research, vol. 67, 2017, pp. 1–13.

[24] P. Rzepakowski and S. Jaroszewicz, "Decision trees for uplift modeling with single and multiple treatments," Knowledge and Information Systems, vol. 32, pp. 303–327, 2012, doi: 10.1007/s10115-011-0434-0.

[25] M. Sołtys, S. Jaroszewicz, and P. Rzepakowski, "Ensemble methods for uplift modeling," Data Mining and Knowledge Discovery, vol. 29, pp. 1531–1559, 2015, doi: 10.1007/s10618-014-0383-9.

[26] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[27] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29–36, 1982, doi: 10.1148/radiology.143.1.7063747.

[28] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, “Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms,” arXiv preprint arXiv:2511.19481, Nov. 2025.

[29] C. Wang, Z. Wen, R. Zhang, P. Xu, and Y. Jiang, “GPU memory requirement prediction for deep learning task based on bidirectional gated recurrent unit optimization transformer,” in Proceedings of the 2025 5th International Conference on Artificial Intelligence, Virtual Reality and Visualization (AIVRV 2025), Oct. 2025.

[30] Z. Wen, R. Zhang, and C. Wang, “Optimization of bi-directional gated loop cell based on multi-head attention mechanism for SSD health state classification model,” in Proceedings of the 2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI). IEEE, 2025, p. 5.

[31] H. Zhang, “LLM-Driven CI Failure Diagnosis and Automated Repair: From GitHub Actions Logs to Patch Recommendation,” Journal of Technology Informatics and Engineering, vol. 4, no. 1, pp. 190–214, Feb. 2025.

[32] 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.

[33] 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.

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

[35] 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.

[36] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, “ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence”, JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[37] 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.

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

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

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

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

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

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

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

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

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

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

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

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

[50] Q. Xin, “Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment”, journalisi, vol. 7, no. 3, pp. 2182-2195, Sep. 2025.

[51] Q. Xin, “Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)”, JTIE, vol. 4, no. 1, pp. 215–238, Feb. 2025, doi: 10.51903/jtie.v4i1.485.

[52] Jubin Zhang, “Graph-based Knowledge Tracing for Personalized MOOC Path Recommendation”, JACS, vol. 5, no. 11, pp. 1–15, Nov. 2025, doi: 10.69987/JACS.2025.51101.

[53] Y. Lu, H. Zhou, and Y. Zhang, “A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling,” Journal of Technology Informatics and Engineering, vol. 4, no. 3, pp. 493–520, Dec. 2025, doi: 10.51903/jtie.v4i3.466.

[54] Meng-Ju Kuo, Boning Zhang, and Maoxi Li, “CryptoFix: Reproducible Detection and Template Repair of Java Crypto API Misuse on a CryptoAPI-Bench–Compatible Benchmark”, JACS, vol. 5, no. 11, pp. 16–33, Nov. 2025, doi: 10.69987/JACS.2025.51102.

[55] Z. S. Zhong, X. Pan, and Q. Lei, “Bridging domains with approximately shared features,” in Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2025.

[56] J. Bai, H. Wang, Q. Wu, and B. Zhang, “Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom E-Mail experiment,” Journal of Technology Informatics and Engineering, vol. 5, no. 1, pp. 17–38, Feb. 2026, doi: 10.51903/jtie.v5i1.468.

[57] Hanqi Zhang, “Prediction Markets as Calibration Teachers for Real-Time Bidding: Market Pricing Meets Ad Auctions”, JACS, vol. 6, no. 1, pp. 1–18, Jan. 2026, doi: 10.69987/JACS.2026.60101.