Uplift Modeling with Delayed Feedback: Identifiability and Algorithms

Anonymous submission

Abstract

Uplift modeling has obtained significant attention, with broad applications in medicine, economics, and marketing. Existing research has typically assumed that one of the potential outcomes of interest can be observed promptly and accurately. Yet in practical settings, treatments always take time to manifest causal impacts on outcomes—for example, the time lag between recommendations or advertisements and consumer purchases. Failing to account for these temporal delays can result in skewed uplift modeling. To address this gap, this work examines how observation timing influences the assessment of uplift by explicitly modeling the potential response time embedded in outcomes. Theoretical analysis establishes the conditions for identifiability under delayed feedback scenarios. We introduce CFR-DF (Counterfactual Regression with Delayed Feedback), a systematic framework that jointly learns both the latent response times and the underlying potential outcomes. Empirical evaluations on synthetic and real-world datasets, including an A/B test with over 1 billion users for 14 days, validate the approach, demonstrating its ability to handle temporal delays and improve estimation accuracy compared to previous uplift modeling methods.

Introduction

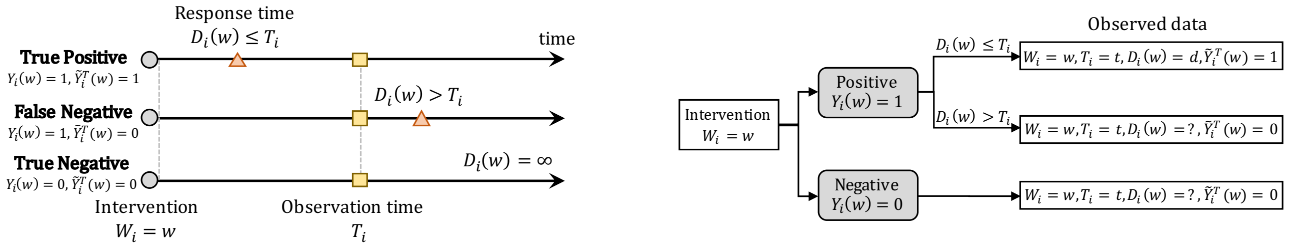
Uplift modeling using observational data is a fundamental problem that applies to a wide variety of areas (Alaa and Van Der Schaar 2017; Alaa, Weisz, and Van Der Schaar 2017; Hannart et al. 2016). For example, in online markets, the uplift of recommending an item (compared to not recommending) on a user's purchase behavior is used for personalized recommendations (Schnabel et al. 2016). Unlike using observed outcomes, Uplift modeling accounts for the difference between factual outcomes and counterfactual outcomes when making decisions. The challenge lies in accurately estimating uplift due to unobserved counterfactual outcomes with alternative treatment (Holland 1986).

Many methods have been proposed to estimate uplift from observational data. For instance, representation learning-based approaches learn a covariate representation that is independent of the treatment to overcome the covariate shift between the treatment and control groups (Johansson, Shalit, and Sontag 2016; Shalit, Johansson, and Sontag 2017; Shi, Blei, and Veitch 2019; Yao et al. 2018). The tree-based approach includes Bayesian inference and random forest methods for nonparametric estimation (Chipman, George, and McCulloch 2010; Wager and Athey 2018). The generative model-based approaches use the widely adopted variational autoencoder and generative adversarial network to generate individual counterfactual outcomes (Louizos et al. 2017; Yoon, Jordon, and Van Der Schaar 2018).

Existing methods require that one of the potential outcomes of interest be observed timely and accurate. However, interventions on individuals usually do not affect outcomes of interest immediately, and treatment takes time to produce causal effects on the outcomes. For example, a recommendation algorithm focuses on whether or not the user will eventually purchase, but users take time to purchase items after being recommended (Chapelle 2014; Yoshikawa and Imai 2018), which poses a critical challenge in practice: as in Figure 1(a), if the observation window is too short, some samples will be incorrectly marked as negative whose conversion will occur in the future. Ignoring such delays in outcome response can lead to biased estimates of uplift.

In this paper, we first formalize the uplift modeling problem in the presence of delayed feedback. In addition to only considering the uplift of treatment on outcome, we also consider different potential response times with different treatments, since treatment may affect response time, e.g., users who receive item recommendations purchase more quickly. Therefore, as in Figure 1(a), given the treatment w for an individual, even the eventual outcome of interest $Y(w)$ is positive, e.g., the user will eventually purchase the item, we can only observe the positive conversion $\tilde{Y}(w) = 1$ when the potential response time is less than the observation time $(D(w) \leq T)$, while observing the false negative outcome $\tilde{Y}(w) = 0$ vise versa. Instead, when the eventual outcome $Y(w)$ is negative, e.g., the user never purchases the item, then we observe the negative outcome $(\tilde{Y}(w) = 0)$ regardless of the observation time. Figure 1(b) illustrates the format of the observed data with an additional challenge compared to the traditional scenario, that is, we could not obtain the exact value of the response time and the true label if the positive feedback did not occur before the observation time.

To address the above issues, we study the impact of observation time on uplift modeling. Theoretically, we prove the eventual potential outcomes are identifiable in the whole population, which is essential for treatment allocation. For subgroups in which individuals always have positive eventual outcomes regardless of treatment, we also show the



Response time
$D_i(w) \le T_i$
time
True Positive
$Y_i(w) = 1, \tilde{Y}_i^T(w) = 1$
False Negative
$D_i(w) > T_i$
$Y_i(w) = 1, \tilde{Y}_i^T(w) = 0$
True Negative
$D_i(w) = \infty$
$Y_i(w) = 0, \tilde{Y}_i^T(w) = 0$
Intervention
$W_i = w$
Observation time
T_i
Observed data
$D_i(w) \le T_i$
$W_i = w, T_i = t, D_i(w) = d, \tilde{Y}_i^T(w) = 1$
Positive
$Y_i(w) = 1$
Intervention
$W_i = w$
$D_i(w) > T_i$
$W_i = w, T_i = t, D_i(w) = ?, \tilde{Y}_i^T(w) = 0$
Negative
$Y_i(w) = 0$
$W_i = w, T_i = t, D_i(w) = ?, \tilde{Y}_i^T(w) = 0$

(a) Three types of delayed feedback scenarios.

(b) Observed data with various potential outcomes.

Figure 1: Illustrations for false negative (left) and data format (right) under delayed feedback.

identifiability of potential response times. Regarding the eventual outcomes as hidden variables, we reconstruct the posterior distribution of delayed feedback and provide explicit solutions to estimate the parameters of interest within a modified EM algorithm. Furthermore, we propose a principled learning approach that extends counterfactual regression to delayed feedback outcomes, named CFR-DF, to simultaneously predict potential outcomes and potential response times. Finally, we validate the effectiveness of the proposed method on both synthetic and real-world datasets. The main contributions are summarized as follows:

- We formalize the uplift modeling problem with delayed feedback, in which treatment takes time to produce a causal effect on the outcome.
- We theoretically prove the eventual potential outcome is identifiable, also show the identifiability of potential response times on the always-positive stratum.
- We propose a principled learning algorithm, called CFR-DF, that utilizes the EM algorithm to estimate both eventual potential outcomes and potential response times.
- We perform extensive experiments on both synthetic and real-world datasets, including an A/B test with over 1 billion users, to show the effectiveness of our approach.

Uplift Modeling with Delayed Feedback

Notation and Setup

In this paper, we consider the case of binary treatment. Suppose we have n units, for each unit i, the covariate and the assigned treatment are denoted as $X_i \in \mathcal{X} \subset \mathbb{R}^m$ and $W_i \in \mathcal{W} = \{0, 1\}$, where $W_i = 1$ means receiving the treatment and $W_i = 0$ means not receiving the treatment, respectively. Compared to the previous uplift modeling methods, we consider the response time from the imposing treatment to producing an influence on the outcome. Specifically, under the potential outcome framework (Rubin 1974; Neyman 1990), let $Y_i(w) \in \mathcal{Y} = \{0, 1\}$ be the binary outcome at the eventual time with treatment w, e.g., whether a user will eventually purchase, as the primary outcome of interest, and we call unit with $Y_i(w) = 1$ as a positive sample. Without loss of generality, the time at which the treatment W_i is imposed on unit i is taken as the start time. Let $D_i(w)$ be the response time for individuals with $Y_i(w) = 1$ to produce positive feedback, and we set $D_i(w) = \infty$ for individuals with $Y_i(w) = 0$. Given an observation time T_i, we see a positive feedback at T_i, denoted as $\tilde{Y}_i^T(w) = 1$, if and only if individual i is a positive sample $Y_i(w) = 1$ with the response time $D_i(w) \le T_i$, and marked as *true positive*. However, we would see false negative feedback $\tilde{Y}_i^T(w) = 0$ at the observation time T_i, when the response time is greater than the observation time, i.e., $D_i(w) > T_i$ with $Y_i(w) = 1$, and marked as *false negative*. For samples that never yield positive outcomes, we observe negative feedback $\tilde{Y}_i^T(w) = 0$ for all observation times T_i, and marked as *true negative*. Since each unit can be only assigned with one treatment, we always observe the corresponding outcome to be either $\tilde{Y}_i^T(0), D_i(0)$ or $\tilde{Y}_i^T(1), D_i(1)$, but not both, which is the fundamental problem of causal inference (Holland 1986; Morgan and Winship 2015).

Parameters of Interest

We consider two meaningful parameters of interest. For simplification, we drop the subscript i hereafter. First, **unlike previous studies that focused on the uplift of treatment on current observed outcomes,** i.e., $\tau^T(x) = \mathbb{E}[\tilde{Y}^T(1) - \tilde{Y}^T(0) \mid X = x]$, **we focused on the uplift on the eventual outcomes,** i.e., $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$. The latter poses two challenges: first, the confounding bias introduced by covariates, which is similar to previous studies; second, how to recover the eventual outcome Y of interest from the observed outcome $\tilde{Y}^T$ at time T.

Next, we show that individuals can be divided into four strata by considering the joint potential outcomes $(Y(0), Y(1))$, as shown in Table 1. From a policy learning perspective, it is clear that treatment should be given to *useful treatment* and not given to *harmful treatment* strata, respectively. For individuals in the *always negative* stratum, either of the treatments is reasonable because the results show no difference. When considering individuals in the *always positive* stratum, despite having both $Y(0) = 1$ and $Y(1) = 1$ for the eventual outcomes, it is meaningful to study the uplift modeling of the treatment on the response times. Formally, the causal estimand of interest is $\mathbb{E}[D(1) - D(0) \mid Y(0) = 1, Y(1) = 1, X = x]$. For the other three strata, since there exists a treatment w such that $Y(w) = 0$, the corresponding response time can be regarded as $D(w) = \infty$, resulting in uplift modeling of treatment on response time being ill-defined.

Group	$Y(0)$	$Y(1)$	$D(0)$	$D(1)$	Preferred treatment
always positive	1	1	✓	✓	Depends on $\tau_D(x)$
useful treatment	0	1	∞	✓	Treatment ($W = 1$)
harmful treatment	1	0	✓	∞	Control ($W = 0$)
always negative	0	0	∞	∞	Either ($W = 0$ or 1)

Table 1: The units are divided into four strata based on the joint potential outcomes $(Y(0), Y(1))$.

We summarize the causal estimand of interest as follows.

- Uplift on the eventual outcome: $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$;
- Uplift on the response time: $\tau_D(x) = \mathbb{E}[D(1) - D(0) \mid Y(0) = 1, Y(1) = 1, X = x]$.

Identifiability Results

We then discuss the identifiability of the causal parameters of interest. Besides some widely used assumptions, such as positivity, consistency, and SUTVA, we adopt the following common assumptions in uplift modeling.

Assumption 1 (Unconfoundedness)

$$W \perp\!\!\!\perp (D(0), D(1), \tilde{Y}^t(0), \tilde{Y}^t(1)) \mid X \quad \textit{for all} \quad t > 0.$$

Assumption 2 (Time Independence)

$$T \perp\!\!\!\perp (D(0), D(1), \tilde{Y}^t(0), \tilde{Y}^t(1), W) \mid X \quad \textit{for all} \quad t > 0.$$

Assumption 3 (Time Sufficiency) $\inf\{d : F_D^{(w)}(d \mid Y(w) = 1, X) = 1\} < \inf\{t : F_T(t) = 1\}$ *for* $w = 0, 1$, *where* $F(\cdot)$ *is the cumulative distribution function (cdf).*

Assumption 4 (Monotonicity) $Y(0) \leq Y(1)$.

Assumption 5 (Principal Ignorability) $(W, Y(w)) \perp\!\!\!\perp D(1 - w) \mid Y(1 - w), X$ *for* $w = 0, 1$.

Among them, unconfoundedness is also known as the no unmeasured confounders assumption, which means all variables that affect both treatment and potential outcomes are included in X. Time independence holds since the observation occurs after the treatment, and the observation does not affect the potential response times $D(w)$ and the potential outcomes $\tilde{Y}^t(w)$ at a given time $t > 0$ for $w = 0, 1$. Time Sufficiency means that we need a subset of individuals (not all) with observed outcomes $\tilde{Y} = 1$ to identify eventual potential outcomes, which is a necessary condition for studying survival analysis. The monotonicity assumption is plausible in many applications when the effect of the decision on the outcome is non-negative for all individuals, e.g., recommendations do not have a negative effect on user purchases. Principal Ignorability requires that the expectations of the potential outcomes do not vary across principal strata conditional on the covariates. It is widely used in applied statistics (Imai and Jiang 2020; Ben-Michael, Imai, and Jiang 2022).

We next provide the identifiability results of three causal parameters (see Appendix for proofs).

Theorem 1 *Under Assumptions 1-3, the uplift on the eventual outcome* $\tau(x)$ *is identifiable.*

In addition, with monotonicity assumption and principal ignorability assumption, we can identify the uplift on potential response times in the *always-positive* stratum $\tau_D(x)$.

Theorem 2 *Under Assumptions 1-5, we can identify the uplift on the response time in the always-positive stratum* $\tau_D(x) = \mathbb{E}[D(1) - D(0) \mid Y(0) = 1, Y(1) = 1, X = x]$.

Note that though assigning treatment on *always-positive* stratum has no effect on $\tau(x)$, if $\tau_D(x)$ is large, it may still be a desirable treatment assignment. For example, even if a customer will buy this commodity, an advertisement may make customers purchase more quickly, enabling merchants to recover costs.

CFR-DF: Counterfactual Regression with Delayed Feedback

In this section, we propose a principled learning approach to perform **C**ounter**F**actual **R**egression with **D**elayed **F**eedback on outcomes, named CFR-DF. Specifically, CFR-DF consists of two sets of models to predict the eventual potential outcomes, i.e., $\mathbb{P}(Y(0) = 1 \mid X = x)$ and $\mathbb{P}(Y(1) = 1 \mid X = x)$ and the potential response times, i.e., $\mathbb{P}(D(0) = d \mid X = x, Y(0) = 1)$ and $\mathbb{P}(D(1) = d \mid X = x, Y(1) = 1)$, respectively, the former of which can be flexibly exploited from previous uplift modeling methods in the following framework, and we take the widely used counterfactual regression (CFR) (Shalit, Johansson, and Sontag 2017) for illustration purpose.

Recall that in Figure 1(b), we show two possible observed data formats. On the one hand, the probability of observing positive feedback $\tilde{Y}^T = 1$ with response time $D = d$ at time $T = t > d$:

$$\begin{aligned}
&p(\tilde{Y}^T = 1, D = d \mid X = x, W = w, T = t) \\
&= p(Y = 1, D = d \mid X = x, W = w) \\
&= \mathbb{P}(Y(w) = 1 \mid X = x, W = w) \\
&\quad \cdot p(D(w) = d \mid X = x, W = w, Y(w) = 1) \\
&= \mathbb{P}(Y(w) = 1 \mid X = x) p(D(w) = d \mid X = x, Y(w) = 1),
\end{aligned}$$

where the first equality follows from time independence, the second equality follows from the consistency assumption, and the last equality follows from the unconfoundedness assumption. To avoid misleading, we use $p(\cdot)$ to represent density, and $\mathbb{P}(\cdot)$ to represent probability.

On the other hand, by the law of total probabilities, and again using the conditional independence of observation time, the probability of not having observed positive feedback at time $T = t > d$ is:

$$\begin{aligned}&\mathbb{P}(\tilde{Y}^T = 0 \mid X = x, W = w, T = t)\\ =&\mathbb{P}(Y = 0 \mid X = x, W = w)\mathbb{P}(\tilde{Y}^t = 0 \mid X = x, W = w, Y = 0)\\ +&\mathbb{P}(Y = 1 \mid X = x, W = w)\mathbb{P}(\tilde{Y}^t = 0 \mid X = x, W = w, Y = 1),\end{aligned}$$

where $\mathbb{P}(Y = 0 \mid X = x, W = w)$ is equivalent to $\mathbb{P}(Y(w) = 0 \mid X = x)$ by unconfoundedness assumption, with similar result holds for $\mathbb{P}(Y = 1 \mid X = x, W = w)$. In addition, we have $\mathbb{P}(\tilde{Y}^t = 0 \mid X = x, W = w, Y = 0) = 1$, due to eventual outcome $Y = 0$ implies $\tilde{Y}^t = 0$ for all $t > 0$. By noting the equivalence between $(\tilde{Y}^t(w) = 0, Y(w) = 1)$ and $(D(w) > t, Y(w) = 1)$:

$$\begin{aligned}&\mathbb{P}(\tilde{Y}^t = 0 \mid X = x, W = w, Y = 1)\\ =\ &\mathbb{P}(D(w) > t \mid X = x, Y(w) = 1)\\ =&\int_t^\infty p(D(w) = u \mid X = x, Y(w) = 1)du.\end{aligned}$$

With the above results, we have the probability of $\tilde{Y}^T = 0$ at time $T = t$ is:

$$\begin{aligned}&\mathbb{P}(\tilde{Y}^T = 0 \mid X = x, W = w, T = t)\\ &= \mathbb{P}(Y(w) = 0 \mid X = x)\\ &+ \mathbb{P}(Y(w) = 1 \mid X = x)\\ &\cdot \int_t^\infty p(D(w) = u \mid X = x, Y(w) = 1)du,\end{aligned}$$

which can be represented by two sets of models in CFR-DF.

Different from CFR, an essential challenge is that we cannot observe the eventual outcomes Y, which results in the unavailability to directly fit the potential outcomes of interest $\mathbb{P}(Y(w) = 0 \mid X = x)$ and $\mathbb{P}(Y(w) = 1 \mid X = x)$ from the observed data. To address this problem, we treat the eventual potential outcomes as latent variables, and estimate the parameters of interest using a modified EM algorithm as below, which addresses both the confounding bias and the missing eventual outcomes.

Expectation Step (E-Step). For a given data point (x_i, w_i, t_i, y_i^t), we need to compute the posterior probability of the hidden variable $p_i := \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i, W = w_i, T = t_i, \tilde{Y}^T = y_i^t)$. If positive feedback $y_i^t = 1$ is observed at time $T = t$, then it is obvious that $p_i = 1$ for unit i. Alternatively, if $y_i^t = 0$ is observed at time t for individual i, then the posterior probability p_i is:

$$\begin{aligned}p_i &= \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i, W = w_i, T = t_i, \tilde{Y}_i^T = 0)\\ &= \frac{\mathbb{P}(\tilde{Y}_i^T(w_i) = 0 \mid X = x_i, Y_i(w_i) = 1, T = t_i)}{\mathbb{P}(\tilde{Y}_i^T = 0 \mid X = x_i, W = w_i, T = t_i)}\\ &\quad \cdot \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i),\end{aligned}$$

which can be calculated from the maximization step of the models in CFR-DR in the following.

Maximization Step (M-Step). Given the hidden variable values p_i computed from the E-step, let $S = s_i$ denote $(X = x_i, W = w_i, T = t_i)$, we maximize the expected log-likelihood in M-step:

$$\begin{aligned}&\sum_i p_i \log \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i)\\ &+ \sum_i (1 - p_i)\log(1 - \mathbb{P}(Y_i(w_i) = 1 \mid X = x_i))\\ &+ \sum_{i:\tilde{y}_i^t = 1} \log p(D_i(w_i) = d_i \mid X = x_i, Y_i(w_i) = 1)\\ &+ \sum_{i:\tilde{y}_i^t = 0} p_i \log \int_{t_i}^\infty p(D(w_i) = u \mid X = x_i, Y_i(w_i) = 1)du,\end{aligned}$$

Then we introduce the representation learning model structure to learn $\mathbb{P}(Y(w) = 1 \mid X = x)$ and $p(D(w) = d \mid X = x, Y(w) = 1)$. Let $h^Y(\Phi^Y(x), w)$ be the prediction model for the eventual potential outcomes $\mathbb{P}(Y(w) = 1 \mid X = x)$, and $h^D(\Phi^D(x), w, d)$ be the prediction model for the potential response times $p(D(w) = d \mid X = x, Y(w) = 1)$, where $\Phi^Y : \mathcal{X} \to \mathcal{R}^Y$ and $\Phi^D : \mathcal{X} \to \mathcal{R}^D$ are the covariate representations, $\mathcal{R}^Y$ and $\mathcal{R}^D$ are the representation spaces, and $h^Y : \mathcal{R}^Y \times \{0, 1\} \to \mathcal{Y}$ and $h^D : \mathcal{R}^D \times \{0, 1\} \times \mathbb{R}^+ \to \mathbb{R}^+$ are the prediction heads, respectively. We take the Integral Probability Metric (IPM) distance induced by the representations as a penalty term, to control the generalization error caused by covariate shift between the treatment and control groups.

Given the posterior probabilities p_i computed from the E-step, we train the eventual potential outcome model by minimizing the derived negative log-likelihood in the M-step with the IPM distance:

$$\begin{aligned}&\ell(h^Y, \Phi^Y \mid p_1, \ldots, p_n) = -\sum_i p_i \log h^Y(\Phi^Y(x_i), w_i)\\ &- \sum_i (1 - p_i)\log(1 - h^Y(\Phi^Y(x_i), w_i))\\ &+ \alpha^Y \cdot \mathrm{IPM}_{\mathcal{G}^Y}(\{\Phi^Y(x_i)\}_{i:w_i=0}, \{\Phi^Y(x_i)\}_{i:w_i=1}),\end{aligned}$$

where $\mathcal{G}^Y$ is a family of functions $g^Y : \mathcal{R}^Y \to \mathcal{Y}$, and α^Y is a hyper-parameter. For two probability density functions p, q defined over $\mathcal{S} \subseteq \mathbb{R}^d$, and for a function family G of functions $g : \mathcal{S} \to \mathbb{R}$, the IPM distance is $\mathrm{IPM}_{\mathrm{G}}(p, q) := \sup_{g \in \mathrm{G}} \left| \int_{\mathcal{S}} g(s)(p(s) - q(s))ds \right|$. Similarly, we train the potential response time model by:

$$\begin{aligned}&\ell(h^D, \Phi^D \mid p_1, \ldots, p_n) = \sum_{i:\tilde{y}_i^t = 1} \log h^D(\Phi^D(x_i), w_i, d_i)\\ &+ \sum_{i:\tilde{y}_i^t = 0} p_i \log \int_{t_i}^\infty h^D(\Phi^D(x_i), w_i, u)du\\ &+ \alpha^D \cdot \mathrm{IPM}_{\mathcal{G}^D}(\{\Phi^D(x_i)\}_{i:w_i=0}, \{\Phi^D(x_i)\}_{i:w_i=1}),\end{aligned}$$

with $\mathcal{G}^D$ and α^D defined similarly. We summarize the algorithm, including the detailed backbone and hyper-parameters choices, in Appendix.

Implementation of CFR-DF. Based on the developed EM algorithm in a general functional form for estimating the parameters of interest, we now show the empirical computation details for computing the integration, i.e., $\mathbb{P}(D(w) >$

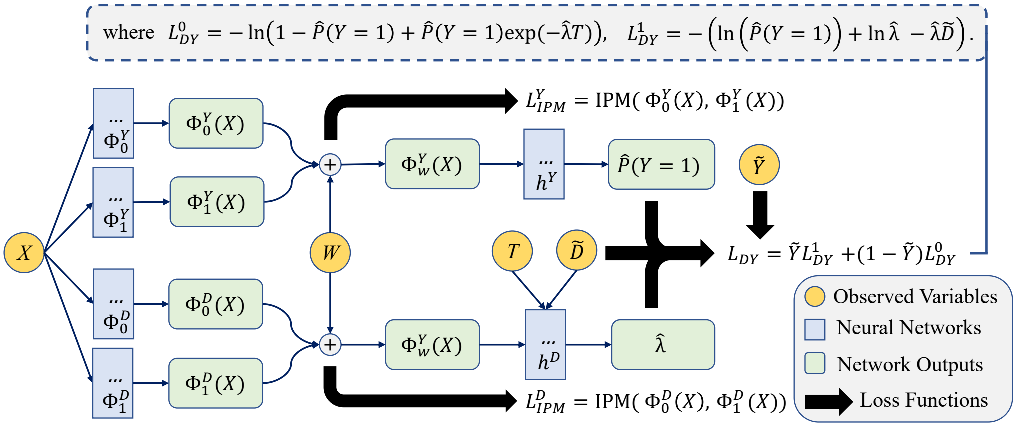

where $L^0_{DY} = -\ln\big(1-\hat{P}(Y=1)+\hat{P}(Y=1)\exp(-\hat{\lambda}T)\big)$, $L^1_{DY} = -\Big(\ln\big(\hat{P}(Y=1)\big)+\ln\hat{\lambda}-\hat{\lambda}\tilde{D}\Big)$.
L^Y_{IPM} = IPM($\Phi^Y_0(X)$, $\Phi^Y_1(X)$)
Φ^Y_0
$\Phi^Y_0(X)$
Φ^Y_1
$\Phi^Y_1(X)$
X
W
$\Phi^Y_w(X)$
h^Y
$\hat{P}(Y=1)$
$\tilde{Y}$
T
$\tilde{D}$
$L_{DY} = \tilde{Y}L^1_{DY} + (1-\tilde{Y})L^0_{DY}$
Φ^D_0
$\Phi^D_0(X)$
Φ^D_1
$\Phi^D_1(X)$
$\Phi^Y_w(X)$
h^D
$\hat{\lambda}$
L^D_{IPM} = IPM($\Phi^D_0(X)$, $\Phi^D_1(X)$)
Observed Variables
Neural Networks
Network Outputs
Loss Functions

Figure 2: Overview of CFR-DF Architecture. For the representation block, we use multi-layer neural networks Φ with ELU activation function to learn representation and each network has two/three layers with m_X units, respectively. Then, we use a single-layer network h^Y with Sigmoid activation to achieve $\hat{P}(Y=1)$ and a single-layer network h^D with SoftPlus sigmoid activation to achieve $\hat{\lambda}$.

$t \mid X = x, Y(w) = 1)$. We here introduce the parametric model method, and we will introduce the non-parametric model based on weighted kernel functions in the Appendix. For a parametric model, One can assume that the potential delayed feedback times obey exponential models for both treatment and control groups. Specifically, let $\mathbb{P}(D(w) = u \mid X = \mathbf{x}, Y(w) = 1) = \lambda_w(\mathbf{x})\exp(-\lambda_w(\mathbf{x})u)$ for $w = 0, 1$, we have $\int_t^\infty \mathbb{P}(D(w) = u \mid X = \mathbf{x}, Y(w) = 1)du = \int_t^\infty \lambda_w(\mathbf{x})\exp(-\lambda_w(\mathbf{x})u)\,du = \exp(-\lambda_w(\mathbf{x})t)$ in the derived p_i in the E-step.

Scalability to Non-Binary Treatments. Our work can be naturally extended to non-binary treatments with the identifiability results of true uplift modeling in all strata, i.e., $\mathbb{E}[Y(w) \mid X = x]$ for all $w \in \mathcal{W}$. By defining delayed feedback time $D(w)$ for all $w \in \mathcal{W}$ similarly and following a similar argument of our identifiability proof, and substitute $Y(0)$ and $Y(1)$ to $Y(w)$ for all $w \in \mathcal{W}$, the true uplift modeling $\mathbb{E}[Y(w) \mid X = x]$ for all $w \in \mathcal{W}$ can be identified similarly. Moreover, in the proposed time-to-event-based uplift modeling problem setup with delayed feedback, the outcome of interest has to be binary to ensure well-definedness. To verify the scalability to non-binary treatments, we conduct an online A/B test over 1 billion users with 13 possible treatments. See the experiment part for a detailed discussion.

Experiments

Baselines and Evaluation Protocols

We evaluate our framework CFR-DF, and its variant without balancing regularization (TAR-DF), in the task of (i) estimating uplift modeling on the eventual outcome and (ii) estimating uplift modeling on the response time in the always-positive stratum. We compare our method with the following methods: **T-learner** (Künzel et al. 2019), representation-based algorithms including **CFR** (Shalit, Johansson, and Sontag 2017), **SITE** (Yao et al. 2018), **Dragonnet** (Shi, Blei, and Veitch 2019), **CFR-ISW** (Hassanpour and Greiner 2019), **DR-CFR** (Hassanpour and Greiner 2020) and **DER-CFR** (Wu et al. 2022), and generative algorithms **CEVAE** (Louizos et al. 2017) and **GANITE** (Yoon, Jordon, and Van Der Schaar 2018). Following the previous studies (Shalit, Johansson, and Sontag 2017; Yao et al. 2018; Wu et al. 2022), we evaluate the performance of uplift modeling using the following two metrics:

$$\epsilon_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}\left((\hat{y}_i(1)-\hat{y}_i(0))-(y_i(1)-y_i(0))\right)^2,$$

$$\epsilon_{\text{ATE}} = \left|\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i(1)-\hat{y}_i(0)-(y_i(1)-y_i(0))\right)\right|,$$

where $\hat{y}_i$ and y_i are predicted and true outcomes. Code is avaliable in the supplementary material.

Datasets

Synthetic Datasets. Since the true potential outcomes are rarely available for real-world, we conduct simulation studies using synthetic datasets as follows. The observed covariates are generated from $X \sim \mathcal{N}(0, I_{m_X})$, where I_{m_X} denotes m_X-degree identity matrix. The observed treatment $W \sim \text{Bern}(\pi(X))$, where $\pi(X) = \mathbb{P}(W = 1 \mid X) = \sigma(\theta_W \cdot X)$, $\theta_W \sim U(-1, 1)$, and $\sigma(\cdot)$ denotes the sigmoid function. For the eventual potential outcomes, we generate the control outcome $Y(0) \sim \text{Bern}(\sigma(\theta_{Y0} \cdot X^2 + 1))$, and the treated outcome $Y(1) \sim \text{Bern}(\sigma(\theta_{Y1} \cdot X^2 + 2))$, where $\theta_{Y0}, \theta_{Y1} \sim U(-1, 1)$. In addition, we generate the potential response time $D(0) \sim \text{Exp}(\exp(\theta_{D0} \cdot X)^{-1})$, and $D(1) \sim \text{Exp}(\exp(\theta_{D1} \cdot X - b_D)^{-1})$, where $\theta_{D0}, \theta_{D1} \sim U(-0.1, 0.1)$, and b_D *controls the heterogeneity of response time functions*. The observation time is generated via $T \sim \text{Exp}(\lambda)$, where λ is the rate parameter of the exponential distribution, and we set $\lambda = 1$ in our experiments, i.e., the average observation time is $\bar{T} = \lambda^{-1} = 1$. Finally, the observed outcome is $\tilde{Y}^T(W) = W \cdot Y(1) \cdot \mathbb{I}(T \geq D(1)) + (1 - W) \cdot Y(0) \cdot \mathbb{I}(T \geq D(0))$, where $\mathbb{I}(\cdot)$ is the indicator function. Based on the data generation process described above, we

	Toy ($b_D = 0$)		**Toy** ($b_D = 0.5$)		**Toy** ($b_D = 1$)	
Method	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}
T-learner	0.535 ± 0.041	0.069 ± 0.024	0.514 ± 0.036	0.028 ± 0.017	0.523 ± 0.028	0.109 ± 0.017
CFR	0.536 ± 0.042	0.071 ± 0.025	0.517 ± 0.037	0.025 ± 0.016	0.523 ± 0.028	0.108 ± 0.016
SITE	0.630 ± 0.058	0.023 ± 0.041	0.646 ± 0.077	0.026 ± 0.020	0.654 ± 0.039	0.128 ± 0.045
Dragonnet	0.612 ± 0.080	0.101 ± 0.055	0.499 ± 0.023	0.028 ± 0.024	0.504 ± 0.018	0.095 ± 0.032
CFR-ISW	0.552 ± 0.057	0.064 ± 0.040	0.602 ± 0.084	0.034 ± 0.024	0.590 ± 0.081	0.122 ± 0.023
DR-CFR	0.539 ± 0.030	0.071 ± 0.032	0.521 ± 0.044	0.032 ± 0.026	0.524 ± 0.038	0.107 ± 0.035
DER-CFR	0.548 ± 0.051	0.051 ± 0.029	0.540 ± 0.037	0.066 ± 0.043	0.568 ± 0.034	0.162 ± 0.032
CEVAE	0.661 ± 0.077	0.123 ± 0.039	0.661 ± 0.077	0.122 ± 0.039	0.661 ± 0.077	0.122 ± 0.039
GANITE	0.672 ± 0.074	0.173 ± 0.037	0.662 ± 0.075	0.147 ± 0.036	0.655 ± 0.076	0.122 ± 0.035
TAR-DF	$\underline{0.416 \pm 0.019}$	$\underline{0.021 \pm 0.008}$	$\underline{0.432 \pm 0.013}$	$\underline{0.017 \pm 0.014}$	$\underline{0.407 \pm 0.016}$	$\underline{0.013 \pm 0.007}$
CFR-DF	**0.409 ± 0.018**	**0.019 ± 0.008**	**0.404 ± 0.014**	**0.013 ± 0.009**	**0.395 ± 0.013**	**0.011 ± 0.009**

Table 2: Performance comparison (MSE ± SD) on synthetic datasets with varying b_D.

Toy ($b_D = 0$)	$\mathbb{P}(D(1) > d \mid Y(0) = 1, Y(1) = 1, X = x) - \mathbb{P}(D(0) > d \mid Y(0) = 1, Y(1) = 1, X = x)$						$\tau_D(x)$
$D > d$	$d = 0.1$	$d = 0.2$	$d = 0.5$	$d = 1.0$	$d = 2.0$	$d = 5.0$	N/A
TAR-DF	0.017 ± 0.003	0.031 ± 0.005	0.056 ± 0.009	0.068 ± 0.012	0.055 ± 0.012	0.015 ± 0.007	0.190 ± 0.030
CFR-DF	**0.014 ± 0.001**	**0.025 ± 0.003**	**0.045 ± 0.005**	**0.054 ± 0.007**	**0.042 ± 0.005**	**0.008 ± 0.002**	**0.152 ± 0.016**
Toy ($b_D = 1$)	$\mathbb{P}(D(1) > d \mid Y(0) = 1, Y(1) = 1, X = x) - \mathbb{P}(D(0) > d \mid Y(0) = 1, Y(1) = 1, X = x)$						$\tau_D(x)$
$D > d$	$d = 0.1$	$d = 0.2$	$d = 0.5$	$d = 1.0$	$d = 2.0$	$d = 5.0$	N/A
TAR-DF	0.025 ± 0.004	0.040 ± 0.007	0.055 ± 0.010	0.054 ± 0.013	0.041 ± 0.014	0.012 ± 0.007	0.321 ± 0.056
CFR-DF	**0.024 ± 0.003**	**0.037 ± 0.005**	**0.048 ± 0.005**	**0.043 ± 0.006**	**0.030 ± 0.006**	**0.006 ± 0.002**	**0.314 ± 0.047**

Table 3: ϵ_{PEHE} of uplift modelings for potential response times with varying b_D.

sample $N = 20,000$ samples for training and $3,000$ samples for testing. We repeat each experiment 10 times to report the mean and standard deviation of the results (ϵ_{PEHE} and ϵ_{ATE}). Moreover, we vary the heterogeneity of response times by setting $b_D \in \{0, 0.5, 1\}$, named the dataset as **Toy** ($b_D = 0$), **Toy** ($b_D = 0.5$), and **Toy** ($b_D = 1$).

Real-World Datasets. We also evaluate our CFR-DF on three widely-adopted real-world datasets: **AIDS** (Hammer et al. 1997; Norcliffe et al. 2023), **Jobs** (LaLonde 1986; Shalit, Johansson, and Sontag 2017), and **Twins** (Almond, Chay, and Lee 2005; Wu et al. 2022). The **AIDS** dataset contains 1,156 patients in 33 AIDS clinical trial units and 7 National Hemophilia Foundation sites in the United States and Puerto Rico. The **Jobs** dataset is built upon randomized controlled trials and aims to assess the effects of job training programs on employment status. The **Twins** dataset is derived from all twins born in the USA between the years 1989 and 1991, and is utilized to assess the influence of birth weight on mortality within one year. For all three datasets, we use the observed covariate X, and following the same procedure for generating synthetic datasets, we generate treatment W, potential outcomes $Y(0)$ and $Y(1)$, potential response times $D(0)$ and $D(1)$, observation time T and factual outcomes $\tilde{Y}^T(W)$. Then we randomly split the samples into training/test with an 8/2 ratio, with 10 repetitions.

Furthermore, we conduct an online A/B test on a real-world recommendation platform with 1 billion users for 14 days with 13 possible treatments, using the proposed method as the experimental group and the baseline without considering delayed feedback as the control group to validate the effectiveness of the proposed method.

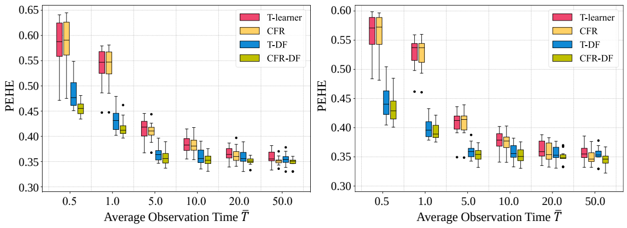
Performance Comparison. We compare our method with the baselines for estimating the uplift in Table 2. The optimal and second-optimal performance are **bold** and underlined. First, the proposed CFR-DF stably outperforms the baselines, as the previous methods do not take into account the delayed feedback, leading to biased estimates of uplift modeling. Second, the TAR-DF method without using balancing regularization slightly degrades the performance compared to CFR-DF, due to the inability to resolve the confounding bias from covariate shift. These results highlight the scalability of our method to varying levels of observation times, showing its potential for real-world applications.

Experiment Results

Table 3 shows the performance of our methods in estimating uplift on the response times. We report the ϵ_{PEHE} on estimating $\mathbb{P}(D(1) > d \mid Y(0) = 1, Y(1) = 1, X = x) - \mathbb{P}(D(0) > d \mid Y(0) = 1, Y(1) = 1, X = x)$ and $\tau_D(x)$, respectively, where the former has a more fine-grained description with varying d. We find that both TAR-DF and CFR-DF can effectively estimate the treatment effect

	AIDS		**Jobs**		**Twins**	
Method	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}	ϵ_{PEHE}	ϵ_{ATE}
T-learner	0.525 ± 0.052	0.091 ± 0.064	0.528 ± 0.043	0.085 ± 0.041	0.390 ± 0.071	0.050 ± 0.029
CFR	0.531 ± 0.046	0.083 ± 0.058	0.510 ± 0.035	0.064 ± 0.039	0.378 ± 0.057	0.029 ± 0.018
SITE	0.601 ± 0.031	0.082 ± 0.056	0.568 ± 0.045	0.064 ± 0.053	0.495 ± 0.087	0.139 ± 0.053
Dragonnet	0.546 ± 0.051	0.105 ± 0.042	0.555 ± 0.060	0.084 ± 0.060	0.440 ± 0.103	0.096 ± 0.067
CFR-ISW	0.592 ± 0.053	0.098 ± 0.032	0.499 ± 0.035	0.058 ± 0.056	0.392 ± 0.048	0.039 ± 0.023
DR-CFR	0.577 ± 0.056	0.078 ± 0.044	0.525 ± 0.077	0.079 ± 0.060	0.390 ± 0.046	0.039 ± 0.027
DER-CFR	0.609 ± 0.076	0.081 ± 0.074	0.503 ± 0.037	0.072 ± 0.043	0.398 ± 0.068	0.080 ± 0.066
CEVAE	0.623 ± 0.042	0.143 ± 0.019	0.638 ± 0.062	0.102 ± 0.058	0.526 ± 0.055	0.139 ± 0.027
GANITE	0.605 ± 0.034	0.136 ± 0.020	0.629 ± 0.053	0.151 ± 0.067	0.509 ± 0.056	0.139 ± 0.040
TAR-DF	0.521 ± 0.042	0.077 ± 0.030	0.453 ± 0.066	0.058 ± 0.030	0.366 ± 0.027	0.030 ± 0.018
CFR-DF	**0.499 ± 0.055**	**0.073 ± 0.031**	**0.438 ± 0.059**	**0.051 ± 0.031**	**0.357 ± 0.017**	**0.027 ± 0.015**

Table 4: Performance comparison (MSE ± SD) on **AIDS**, **Jobs**, and **Twins** datasets.



(a) **Toy** ($b_D = 0$).

(b) **Toy** ($b_D = 1$).

Figure 3: Effects of varying average observation time on synthetic datasets with varying b_D.

on response time, and the CFR-DF with balancing regularization stably performs better, again demonstrating the need to adjust for confounding bias. In Appendix, we further conduct the experiments with various number of features.

Ablation Studies. Figure 3 compares the proposed CFR-DF and its ablated versions for estimating uplift on the eventual outcome with varying average observation time, where TAR-DF does not perform balancing regularization, CFR does not consider delayed feedback, and neither is considered for T-learner. We have the following findings. The proposed CFR-DF and TAR-DF have significantly better performance when the observation time is shorter, due to their effective adjustment for delayed feedback. When increasing the average observation time leads to more delayed feedback being observed, we find improved performance for all four methods. When the observation time reaches 50, meaning almost all delayed feedbacks have been observed, our method performs similarly to the CFR.

Real-World Experiments. We conduct real-world experiments using **AIDS**, **Jobs**, and **Twins** datasets. Notably, treating AIDS requires long-term observation, job training takes time to cause changes in incomes, and infants also take time to observe their mortality outcomes (and thus study the effect on mortality), therefore it is reasonable to study the delayed feedback in such real-world applications. Table 4 demonstrates that CFR-DF outperforms all baselines on these real-world datasets, showcasing its effectiveness.

Metrics	Day3	Day7	Day14
TAU% ↑	**+0.017 ± 0.011**	**+0.019 ± 0.011**	**+0.017 ± 0.012**
CAU% ↑	**+0.015 ± 0.007**	**+0.016 ± 0.007**	**+0.018 ± 0.007**
TCU% ↓	**-0.002 ± 0.720**	**-0.333 ± 0.756**	**-0.600 ± 0.741**
CCU% ↓	**-0.152 ± 0.400**	**-0.255 ± 0.314**	**-0.259 ± 0.239**

Table 5: Results of the **Online A/B Test.** Metrics with an upward arrow indicate that higher values are preferable. Compared to the baseline, we report the results as **average difference ± confidence interval width.**

Online A/B Test. Table 5 shows online A/B test results on a real-world platform, comparing our proposed method against the baseline without considering delayed feedback, using DESCN (Zhong et al. 2022) as the backbone. We evaluate performance with four metrics: Today active user (TAU), Today close user (TCU), Cumulated active user (CAU), and Cumulated close user (CCU). Overall, our method increases active users by 0.0176% and reduces close users by 0.259%, demonstrating its effectiveness in a non-binary treatment industry scenario. Moreover, these results also show that our method is backbone agnostic.

Conclusion

This paper studies the uplift modeling problem by further considering the response time needed for a treatment to produce a causal effect on the outcome. Specifically, we propose a principled learning algorithm, called CFR-DF, to estimate both eventual potential outcomes and potential response times. Considering the widespread of delayed feedback outcomes, we believe such study is meaningful for real-world applications. A shortcoming of our study is the validity of the assumptions in practice, e.g., we need enough observation time to identify uplift modeling on the eventual potential outcome, and principal ignorability is further required to identify uplift modeling on the response time.

References

Alaa, A. M.; and Van Der Schaar, M. 2017. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30.

Alaa, A. M.; Weisz, M.; and Van Der Schaar, M. 2017. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*.

Almond, D.; Chay, K. Y.; and Lee, D. S. 2005. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3): 1031–1083.

Ben-Michael, E.; Imai, K.; and Jiang, Z. 2022. Policy learning with asymmetric utilities. *arXiv preprint arXiv:2206.10479*.

Chapelle, O. 2014. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1097–1105.

Chipman, H. A.; George, E. I.; and McCulloch, R. E. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1): 266–298.

Hammer, S. M.; Squires, K. E.; Hughes, M. D.; Grimes, J. M.; Demeter, L. M.; Currier, J. S.; Eron Jr, J. J.; Feinberg, J. E.; Balfour Jr, H. H.; Deyton, L. R.; et al. 1997. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine*, 337(11): 725–733.

Hannart, A.; Pearl, J.; Otto, F.; Naveau, P.; and Ghil, M. 2016. Causal counterfactual theory for the attribution of weather and climate-related events. *Bulletin of the American Meteorological Society*, 97(1): 99–110.

Hassanpour, N.; and Greiner, R. 2019. CounterFactual Regression with Importance Sampling Weights. In *IJCAI*, 5880–5887.

Hassanpour, N.; and Greiner, R. 2020. Learning Disentangled Representations for CounterFactual Regression. In *International Conference on Learning Representations*.

Holland, P. W. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81: 945–960.

Imai, K.; and Jiang, Z. 2020. Principal fairness for human and algorithmic decision-making. *arXiv preprint arXiv:2005.10400*.

Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, 3020–3029. PMLR.

Künzel, S. R.; Sekhon, J. S.; Bickel, P. J.; and Yu, B. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10): 4156–4165.

LaLonde, R. J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604–620.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Morgan, S. L.; and Winship, C. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, second edition.

Neyman, J. S. 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5: 465–472.

Norcliffe, A.; Cebere, B.; Imrie, F.; Lio, P.; and van der Schaar, M. 2023. SurvivalGAN: Generating Time-to-Event Data for Survival Analysis. In *International Conference on Artificial Intelligence and Statistics*, 10279–10304. PMLR.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66: 688–701.

Schnabel, T.; Swaminathan, A.; Singh, A.; Chandak, N.; and Joachims, T. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *ICML*.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 3076–3085. PMLR.

Shi, C.; Blei, D.; and Veitch, V. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

Wager, S.; and Athey, S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242.

Wu, A.; Yuan, J.; Kuang, K.; Li, B.; Wu, R.; Zhu, Q.; Zhuang, Y.; and Wu, F. 2022. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4989–5001.

Yao, L.; Li, S.; Li, Y.; Huai, M.; Gao, J.; and Zhang, A. 2018. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31.

Yoon, J.; Jordon, J.; and Van Der Schaar, M. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Yoshikawa, Y.; and Imai, Y. 2018. A nonparametric delayed feedback model for conversion rate prediction. *arXiv preprint arXiv:1802.00255*.

Zhong, K.; Xiao, F.; Ren, Y.; Liang, Y.; Yao, W.; Yang, X.; and Cen, L. 2022. Descn: Deep entire space cross networks for individual treatment effect estimation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 4612–4620.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) yes

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) yes

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) yes

2.4. Proofs of all novel claims are included (yes/partial/no) yes

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) yes

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) yes

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) yes

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) yes

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) partial

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) partial

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) yes

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) partial

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/-partial/no) partial

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/-partial/no) partial

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) partial