Imbalance-Aware Uplift Modeling for Observational Data

Xuanying Chen*[1], Zhining Liu*[1], Li Yu*[1], Liuyi Yao*[2], Wenpeng Zhang[1], Yi Dong[1], Lihong Gu[1], Xiaodong Zeng[1], Yize Tan[1], Jinjie Gu[1]

[1]Ant Group, [2]Alibaba Group
{xuanying.cxy,eason.lzn,jinli.yl}@antgroup.com, yly287738@alibaba-inc.com, zhangwenpeng0@gmail.com,
{dongyi.dy,lihong.glh,xiaodong.zxd,yize.tyz,jinjie.gujj}@antgroup.com

Abstract

Uplift modeling aims to model the incremental impact of a treatment on an individual outcome, which has attracted great interests of researchers and practitioners from different communities. Existing uplift modeling methods rely on either the data collected from randomized controlled trials (RCTs) or the observational data which is more realistic. However, we notice that on the observational data, it is often the case that only a small number of subjects receive treatment, but finally infer the uplift on a much large group of subjects. Such highly imbalanced data is common in various fields such as marketing and medical treatment but it is rarely handled by existing works. In this paper, we theoretically and quantitatively prove that the existing representative methods, transformed outcome (TOM) and doubly robust (DR), suffer from large bias and deviation on highly imbalanced datasets with skewed propensity scores, mainly because they are proportional to the reciprocal of the propensity score. To reduce the bias and deviation of uplift modeling with an imbalanced dataset, we propose an imbalance-aware uplift modeling (IAUM) method via constructing a robust proxy outcome, which adaptively combines the doubly robust estimator and the imputed treatment effects based on the propensity score. We theoretically prove that IAUM can obtain a better bias-variance trade-off than existing methods on a highly imbalanced dataset. We conduct extensive experiments on a synthetic dataset and two real-world datasets, and the experimental results well demonstrate the superiority of our method over state-of-the-art.

Introduction

Uplift modeling refers to the techniques that model the *incremental impact* of a treatment on an individual outcome, and the incremental impact is also known as *individual treatment effect* (ITE) or the *uplift*. Uplift modeling is widely applied in various domains, such as marketing (Radcliffe 2007), social science (Imai and Ratkovic 2013; Künzel et al. 2019) and medicine treatment (Zhang et al. 2017) because of its ability to sufficiently target customer. For example, in the marketing area, uplift modeling helps the marketing team improve the targeting by focusing on only the persuadable customers who will purchase if they are exposed to a campaign otherwise not. In this way, the no effect or negative effect of a campaign can be prevented, so that the return of investment of a campaign can be maximized.

Traditional uplift modeling methods rely on data collected through RCTs, where subjects are randomly assigned to receive treatment. However, due to the high cost, time-consuming, and sometimes unethical of conducting RCT, a more realistic way is to build the estimator from non-random data, namely observational data. Existing works, such as TOM (Athey and Imbens 2015), DR (Wang et al. 2019) and SDRM (Saito, Sakata, and Nakata 2019) transforms the estimated ITE as the proxy outcome, which can train a new model with any existing off-the-shelf supervised methods to estimate the uplift directly.

However, it is worthy to notice that, existing works may still have large bias and deviation in the highly imbalanced dataset. A highly imbalanced dataset refers to the case where only a small portion of the people receive the treatment, which is very common in various fields. For example, in marketing compaigns, to maximize the return of investment, the advertisement is usually exposed to a small group of audiences to save the cost. In other words, in the observed dataset, only a small group of people receive the treatment, but the estimated uplift model is used to predict the uplift over all the audiences. In this case, particularly small propensity scores exist in the estimation, leading to large estimation bias and deviation in the existing work when its inverse is adopted to construct the proxy outcome, such as TOM and DR. Theoretical and quantitative analysis on TOM and DR show that both methods will suffer the same large bias and deviation problem while learning with a highly imbalanced dataset.

In addition to the extreme propensity score, the certainty level of treated and control outcome prediction would be different due to the high imbalance in treatment and control group size. When constructing the proxy outcome, without considering this certainty difference, the performance of the uplift modeling would be decreased. The above challenges brought by high imbalance data require the estimator to carefully take imbalance into consideration and balance the bias and deviation of the estimator on the two groups with dramatically different numbers of samples.

To overcome the above challenges, we propose an **I**mbalance-**A**ware **U**plift **M**odeling (**IAUM**) by adaptively taking the advantages of the doubly robust estimator and

*These authors contributed equally to this work.

the imputed treatment effects. Specifically, we construct the proxy outcome by aggregating the two estimators weighted by the propensity score because, to some degree, the uncertainty level of the doubly robust estimator and the imputed treatment effects are correlated with the propensity score in this high-imbalance case. Moreover, utilizing the propensity score as the weight avoids its appearance in the denominator compared with the existing work. Furthermore, we theoretically prove that IAUM can obtain a better bias-variance trade-off than existing methods on a highly imbalanced dataset. One thing to note is that, in this paper, we only discuss the case that the treatment probability is relatively small, and the case that the non-treated probability is much smaller than the treated probability can be analyzed in the same way. We conduct the proposed IAUM method on both synthetic and real-world datasets, and the experimental results confirm the effectiveness. In summary, our main contributions are as follows:

- **Problem.** We theoretically and quantitatively prove that the TOM and DR suffer from large bias and deviation on a highly imbalanced dataset with an extreme propensity score, and identify its unique challenges arising from real applications.
- **Method.** To reduce the bias and deviation of uplift modeling with a imbalanced dataset, we propose an imbalance-aware uplift modeling method via constructing a robust proxy result and obtain a better bias-variance trade-off than existing methods.
- **Evaluation.** We perform extensive experiments on a synthetic dataset with eight different scenarios and two real-world datasets, which demonstrates that the proposed method achieves consistent improvement over existing uplift modeling methods.

Related Work

As a method of obtaining ITE, there is a growing interest (Gutierrez and Gérardy 2017; Zhang, Li, and Liu 2020; Yao et al. 2020; Olaya, Coussement, and Verbeke 2020) in developing a unbiased and robust estimator for uplift modeling.

When the RCT data is enough to estimate the ITE, the single model approach (Lo 2002), which uses the concatenation of treatment and covariates to predict the outcomes, and TMA (Jaskowski and Jaroszewicz 2012), which defines ITE as the difference between predicted outcomes coming from two group of subjects, can be directly applied to estimate the uplift.

As it is often difficult to collect the RCT data, a realistic to model the uplift is based on observational data (Nichols 2007). TOM (Athey and Imbens 2015) is one of the most representative method. It uses an unbiased estimator of ITE as a proxy outcome, but requires the propensity score to be unbiased. Since the propensity score is difficult to predict accurately, TOM has been suffering from bias and excessive variance. In order to solve this problem, Wang et al. (2019) presented a doubly robust (DR) technique that combines error imputation based estimator and inverse propensity score estimator. Moreover, Saito, Sakata, and Nakata (2019) introduced a switching approach that switches between DR estimator and predicted treatment effects, which achieves a desirable bias-variance trade-off.

Methods mentioned above usually do not restrict to one specific machine learning approach, and there are another line of research work focusing on reforming the traditional machining learning methods for uplift modeling. Based on binary tree models, Hansotia and Rukstales (2002) proposed a new splitting criterion that maximizes the difference between the estimated treatment effect of the two child nodes. Following the idea of support vector machine (SVM), (Zaniewicz and Jaroszewicz 2013) presented two SVM-based uplift modeling methods, which are the L_1-Uplift Support Vector Machine and the L_p Uplift Support Vector Machine.

Besides the above tailored uplift modeling approach based on traditional machine learning methods, researchers also explore to apply the deep learning techniques to the uplift modeling (Gutierrez and Gérardy 2017). The main advantages of deep learning methods are that the large model capacity of neural networks can easily model complex nonlinear relationships between the treatment and the covariates. In addition, with the flexibility of the design of neural networks, it is easy to realize deconfounding of the uplift modeling on the non-RCT data. Several deep learning based methods (Johansson, Shalit, and Sontag 2016; Yao et al. 2018; Yu et al. 2021; Ma, Li, and Cottrell 2020; Li et al. 2021; Künzel et al. 2018; Yao et al. 2019; Zhang, Liu, and Li 2020; Chen et al. 2021; Yao et al. 2021) successfully extend the traditional approach to combine with deep learning and achieve improvements on the uplift modeling.

However, the aforementioned methods have not discussed the case that learning with a highly imbalanced dataset, and it is still an open question that how to obtain a reliable estimator under this setting. In this paper, we take two representative methods, TOM and DR, as examples for detailed analysis and develop a robust uplift estimator.

Preliminaries

In this section, we first introduce notations according to the Rubin Causality Model (Imbens and Rubin 2015), then introduce the most relevant works TOM (Athey and Imbens 2015) and DR (Wang et al. 2019) in detail.

Notations

The number of users contained in the sample is N, and we use $\mathbf{X}_i \in \mathcal{X}$ to represent the feature vector embedding of i-th user u_i. We denote $W_i \in \mathcal{T}$ as a binary indicator that represents u_i's treatment assignment, i.e.,

$$W_i = \begin{cases} 1, & \text{if } u_i \text{ receives the treatment;} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Let Y_i^1 denote u_i's outcome when they receive the treatment, and Y_i^0 denote u_i's outcome when they do not receive the treatment. Then, we can define our interest scalar τ_i which represents ITE as the difference between two variables:

$$\tau_i = \mathbb{E}[Y_i^1 - Y_i^0 | \mathbf{X}_i]. \tag{2}$$

However, in real life, for any individual, we can only observe one potential result. We represent u_i's observed outcome Y_i^{obs} as:

$$Y_i^{\text{obs}} = W_i Y_i^1 + (1 - W_i) Y_i^0. \tag{3}$$

In addition, we use μ_i^1 and μ_i^0 to represent the user's expectation of potential outcomes conditioned on $\mathbf{X}_i$, which means that $\mu_i^1 = E[Y_i^1|\mathbf{X}_i]$ and $\mu_i^0 = E[Y_i^0|\mathbf{X}_i]$. $\hat{\mu}_i^1$ and $\hat{\mu}_i^0$ are the predicted values of model for μ_i^1 and μ_i^0, respectively. We use $e(\mathbf{X}_i)$ be the propensity score that represents the probability of u_i's being treated, which is written as:

$$e(\mathbf{X}_i) = P(W_i = 1|\mathbf{X}_i) = \mathbb{E}[W_i|\mathbf{X}_i]. \tag{4}$$

We define δ_i^1, δ_i^0 as the deviations between the expected true value for theoretical analysis and the predicted value of the model output, and Δ_i^1, Δ_i^0 represent the deviation of u_i's outcome and the predicted value of the model i.e.,

$$\delta_i^1 = \hat{\mu}_i^1 - \mu_i^1, \delta_i^0 = \hat{\mu}_i^0 - \mu_i^0; \tag{5}$$

$$\Delta_i^1 = Y_i^1 - \hat{\mu}_i^1, \Delta_i^0 = Y_i^0 - \hat{\mu}_i^0. \tag{6}$$

Transformed Outcome Method

Since we cannot observe the real ITE, TOM (Athey and Imbens 2015) uses inverse propensity score to construct an outcome as a proxy for ITE, which is defined as:

$$\begin{aligned} Y_i^{\text{TOM}} =& Y_i^{\text{obs}} \frac{W_i - e(\mathbf{X}_i)}{e(\mathbf{X}_i)(1 - e(\mathbf{X}_i))} \\ =& \begin{cases} \dfrac{Y_i^1}{e(\mathbf{X}_i)}, & W_i = 1; \\ \dfrac{-Y_i^0}{1 - e(\mathbf{X}_i)}, & W_i = 0. \end{cases} \end{aligned} \tag{7}$$

With the above transformed outcome, any off-the-shelf supervised methods can be directly applied for the estimation on the dataset. However, the condition for using TOM to construct an unbiased proxy value is to obtain the true propensity score of each individual (that is $e(\mathbf{X}_i)$), but in practice, the true propensity score cannot be estimated due to the complex data distribution.

Doubly Robust Method

As proved by Saito, Sakata, and Nakata (2019), TOM will be an unreliable proxy outcome with a biased propensity score, therefore, the doubly robust method proposed by (Wang et al. 2019) can be adopted to construct a new proxy outcome with better bias and variance, which is defined as:

$$\begin{aligned} \hat{Y}_i^{\text{DR}} =& \big(\hat{\mu}_i^1 + \frac{W_i}{\hat{e}(\mathbf{X}_i)}(Y_i^{obs} - \hat{\mu}_i^1)\big) \\ & - \big(\hat{\mu}_i^0 + \frac{1 - W_i}{1 - \hat{e}(\mathbf{X}_i)}(Y_i^{obs} - \hat{\mu}_i^0)\big) \\ =& \begin{cases} \hat{\mu}_i^1 - \hat{\mu}_i^0 + \dfrac{Y_i^1 - \hat{\mu}_i^1}{\hat{e}(\mathbf{X}_i)}, & W_i = 1; \\ \hat{\mu}_i^1 - \hat{\mu}_i^0 - \dfrac{Y_i^0 - \hat{\mu}_i^0}{1 - \hat{e}(\mathbf{X}_i)}, & W_i = 0, \end{cases} \end{aligned} \tag{8}$$

where $\hat{e}(\mathbf{X}_i)$ is the estimated value of the propensity score $e(\mathbf{X}_i)$.

Methodology

In this section, we first introduce the challenges of uplift modeling in highly imbalanced dataset. Then we elaborate the proposed imbalance-aware uplift modeling approach. At last, the theoretical analysis about the bias and the deviation of the proposed method is provided.

Challenges

As mentioned previously, to optimize the allocation of the budget, it is common that the treatment is only exposed to a small portion of group because of the limited budget, which makes the size of the treatment and the control group highly imbalanced. Such imbalance leads to two major challenges in uplift modeling: (1) high bias and deviation (2) different difficulty level of treatment/control outcome prediction.

Challenge 1: High Bias and Deviation. Due to the extremely small treatment group size, the estimated propensity scores $\hat{e}(\mathbf{X}_i)$ on some units tend to be very close to 0. Once the estimated propensity score appears in the denominator, as in TOM (Equation (7)) and DR (Equation (8)), it causes high bias and deviation in the highly imbalanced dataset. Formally, the following two theorems shows the bias and deviation of the state-of-the-art methods TOM and DR (See Appendix for the proof).

Theorem 1. *The bias of TOM estimator is*

$$\text{Bias}(\hat{Y}_i^{\text{TOM}}|\mathbf{X}_i) = \begin{cases} \big|\mu_i^0 + \dfrac{1 - \hat{e}(\mathbf{X}_i)}{\hat{e}(\mathbf{X}_i)}\mu_i^1\big|, & W_i = 1; \\ \big|\dfrac{\hat{e}(\mathbf{X}_i)}{1 - \hat{e}(\mathbf{X}_i)}\mu_i^0 + \mu_i^1\big|, & W_i = 0. \end{cases} \tag{9}$$

The deviation of TOM estimator is

$$\Lambda^{\text{TOM}} = \sqrt{C \sum_{i=1}^{n} \big(\frac{Y_i^1}{\hat{e}(\mathbf{X}_i)} + \frac{Y_i^0}{1 - \hat{e}(\mathbf{X}_i)}\big)^2}, \tag{10}$$

where C is a constant.

Theorem 2. *The bias of DR estimator is:*

$$\text{Bias}(\hat{Y}_i^{\text{DR}}|\mathbf{X}_i) = \begin{cases} \big|\delta_i^0 + \dfrac{1 - \hat{e}(\mathbf{X}_i)}{\hat{e}(\mathbf{X}_i)}\delta_i^1\big|, & W_i = 1; \\ \big|\dfrac{\hat{e}(\mathbf{X}_i)}{1 - \hat{e}(\mathbf{X}_i)}\delta_i^0 + \delta_i^1\big|, & W_i = 0. \end{cases} \tag{11}$$

The deviation of DR estimator is:

$$\Lambda^{\text{DR}} = \sqrt{C \sum_{i=1}^{n} \big(\frac{\Delta_i^1}{\hat{e}(\mathbf{X}_i)} + \frac{\Delta_i^0}{1 - \hat{e}(\mathbf{X}_i)}\big)^2}. \tag{12}$$

The above theorems show that the bias and deviation of TOM and DR are proportional to the reciprocal of the propensity score, thus also gives a large bias and deviation. Although DR improved the TOM in reducing the bias and variance, it still lacks the capacity to handle the case where the group sizes of control and treatment group are extremely imbalanced. If $\hat{e}(\mathbf{X}_i) = 0.1$, then $\frac{1}{\hat{e}(\mathbf{X}_i)} = 10$, which gives quite large and unreliable proxy outcome when $W_i = 1$.

Challenge 2: Different Difficulty Level of Outcome Prediction. Since the group size of the control group is much larger than the treatment group, it is more difficult for the predictor to estimate the potential the treatment outcome μ_i^1 than control outcome μ_i^0. In other words, the predicted control outcome $\hat{\mu}_i^0$ are more accurate than the predicted treatment outcome $\hat{\mu}_i^1$. This challenge motivates us that in designing the transformed outcome, the part that contains the predicted control outcome can assign more weights than the part that contains the predicted treatment outcome.

Imbalance-Aware Uplift Modeling

Motivation. The two challenges mentioned above motivate our transformed outcome design that it is a need to prevent the propensity score appearing in the denominator and meanwhile, the uncertainty difference of $\hat{\mu}_i^0$ and $\hat{\mu}_i^1$ should be taken into account.

Proxy Outcome Construction. To solve the above challenges, we proposed a novel method named imbalance-aware uplift model (IAUM), which adaptively combines the doubly robust estimator and the imputed treatment effects based on the propensity score to reduce the bias. The proposed IAUM method is defined as:

$$\begin{aligned}\hat{Y}_i^{\text{IAUM}} &= \begin{cases} \hat{e}(\mathbf{X}_i) * \hat{Y}_i^{\text{DR}} \\ \quad + (1-\hat{e}(\mathbf{X}_i)) * (Y_i^1 - \hat{\mu}_i^0), \quad W_i = 1; \\ (1-\hat{e}(\mathbf{X}_i)) * \hat{Y}_i^{\text{DR}} \\ \quad + \hat{e}(\mathbf{X}_i) * (\hat{\mu}_i^1 - Y_i^0), \quad W_i = 0, \end{cases} \\ &= \begin{cases} Y_i^1 - \hat{\mu}_i^0 + \big(1-\hat{e}(\mathbf{X}_i)\big)(Y_i^1 - \hat{\mu}_i^1), \quad W_i = 1; \\ \hat{\mu}_i^1 - Y_i^0 + \hat{e}(\mathbf{X}_i)(\hat{\mu}_i^0 - Y_i^0), \quad W_i = 0. \end{cases} \end{aligned} \tag{13}$$

where $\hat{Y}_i^{\text{DR}}$ is the doubly robust estimator which is defined as:

$$\begin{aligned}\hat{Y}_i^{\text{DR}} =& \frac{W_i}{\hat{e}(\mathbf{X}_i)}\left(Y_i^{\text{obs}} - \hat{\mu}_i^1\right) \\ & - \frac{1-W_i}{1-\hat{e}(\mathbf{X}_i)}\left(Y_i^{\text{obs}} - \hat{\mu}_i^0\right) + \left(\hat{\mu}_i^1 - \hat{\mu}_i^0\right). \end{aligned} \tag{14}$$

From Equation (13), the doubly robust estimator $\hat{Y}_i^{\text{DR}}$ and the imputed treatment effect are aggregated with the estimated propensity score as their weights. By multiplying estimated propensity score $\hat{e}(\mathbf{X}_i)$ with $\hat{Y}_i^{\text{DR}}$ when $W_i = 1$, the $\hat{e}(\mathbf{X}_i)$ in the denominator of $\hat{Y}_i^{\text{DR}}$ can be cancelled. Furthermore, IAUM fully utilizes the samples collected from the control group and put a small weight on the $\hat{Y}_i^{\text{DR}}$ because it would have large variation due to the small value of $\hat{e}(\mathbf{X}_i)$. When $W_i = 0$, a large weight will put on the $\hat{Y}_i^{\text{DR}}$ since $\hat{\mu}_i^1$ is quite difficult to estimate precisely due to the insufficient samples.

Implementation. To estimate the uplift using IAUM, we first need to build two separate estimators, $\hat{\mu}_i^1$ and $\hat{\mu}_i^0$, using the observed outcome of subjects from the treated group and control group. Then based on how the treatments are distributed to subjects, we estimate the propensity score function $\hat{e}(\mathbf{X}_i)$. With the estimated $\hat{\mu}_i^0$, $\hat{\mu}_i^1$ and $\hat{e}(\mathbf{X}_i)$, we can construct the proxy outcome $\hat{Y}_i^{\text{IAUM}}$ using Equation (13). Finally, any machine learning model can be used to fit $\hat{Y}_i^{\text{IAUM}}$ using $\mathbf{X}_i$ and gives the uplift model. We summarize implementation details of the proposed method in Algorithm 1.

Algorithm 1: IAUM Method

Input: Training data: $D = \{(\mathbf{X}_i, W_i, Y_i^{\text{obs}})\}_{i=1}^N$
Output: Fitted uplift estimator p

1: Fit g to the potential outcome μ_i^0 of the control group using the data $\{(\mathbf{X}_i, Y_i^0)\}_{i=1}^{N_0}$.
2: Fit h to the potential outcome μ_i^1 of the treatment group using the data $\{(\mathbf{X}_i, Y_i^1)\}_{i=1}^{N_1}$.
3: Fit f to estimate propensity score $\hat{e}(\mathbf{X}_i)$ using the data $\{(\mathbf{X}_i, W_i)\}_{i=1}^N$.
4: With estimated g, h and f, construct the proxy outcome Y_i^{IAUM} using Equation (13).
5: Fit p to the data $\{(\mathbf{X}_i, Y_i^{\text{IAUM}})\}_{i=1}^N$.
6: **return** p

Bias Analysis of IAUM

In this subsection, we compare the bias of IAUM with TOM and DR to validate its superiority. Firstly, we prove that $\hat{Y}_i^{\text{IAUM}}$ has the following bias with a biased propensity score estimator.

Theorem 3. *The bias of IAUM estimator is*

$$\text{Bias}(\hat{Y}_i^{\text{IAUM}}|\mathbf{X}_i) = \begin{cases} \left|\delta_i^0 + \big(1-\hat{e}(\mathbf{X}_i)\big)\delta_i^1\right|, \quad W_i = 1; \\ \left|\hat{e}(\mathbf{X}_i)\delta_i^0 + \delta_i^1\right|, \quad W_i = 0. \end{cases} \tag{15}$$

Proof. Given the definition of IAUM estimator in Equation (13), the bias of IAUM is:

$$\begin{aligned} & \text{Bias}(\hat{Y}_i^{\text{IAUM}}|\mathbf{X}_i, W_i = 1) \\ =& \left|\mathbb{E}[\hat{Y}_i^{\text{IAUM}}|\mathbf{X}_i, W_i = 1] - \tau_i\right| \\ =& \left|\mathbb{E}[Y_i^1 - \hat{\mu}_i^0 + \big(1-\hat{e}(\mathbf{X}_i)\big)(Y_i^1 - \hat{\mu}_i^1)|\mathbf{X}_i] - \tau_i\right| \\ =& \left|\mu_i^1 - \hat{\mu}_i^0 + \big(1-\hat{e}(\mathbf{X}_i)\big)(\mu_i^1 - \hat{\mu}_i^1) - (\mu_i^1 - \mu_i^0)\right| \\ =& \left|\big(1-\hat{e}(\mathbf{X}_i)\big)(\hat{\mu}_i^1 - \mu_i^1) + (\hat{\mu}_i^0 - \mu_i^0)\right| \\ =& \left|\delta_i^0 + \big(1-\hat{e}(\mathbf{X}_i)\big)\delta_i^1\right|, \end{aligned} \tag{16}$$

where $\delta_i^0 = \hat{\mu}_i^0 - \mu_i^0$ and $\delta_i^1 = \hat{\mu}_i^1 - \mu_i^1$, and line 3 to line 4 in the above equation is because $\mu_i^1 = E[Y_i^1|\mathbf{X}_i]$ and $\mu_i^0 = E[Y_i^0|\mathbf{X}_i]$. Similarly, when $W_i = 0$, the bias of $\hat{Y}_i^{\text{IAUM}}$ can be derived in the same way.

□

Proposition 1. *Suppose* $|\delta_i^0| < |\delta_i^1|$, *the bias of proposed IAUM method is less than DR and TOM, i.e.,* $\text{Bias}(\hat{Y}_i^{\text{IAUM}}|\mathbf{X}_i, W_i) < \text{Bias}(\hat{Y}_i^{\text{DR}}|\mathbf{X}_i, W_i) < \text{Bias}(\hat{Y}_i^{\text{TOM}}|\mathbf{X}_i, W_i)$.

Proof. According to the above assumptions, the treatment deviations δ_i^1 between the expected true value and the predicted value is much large than δ_i^0, due to lack of treatment samples, i.e. $|\frac{\delta_i^0}{\delta_i^1}| \ll 1$. Under this condition, we can derive the difference of squared bias of IAUM and DR as:

$$\begin{aligned}
\triangle =& \mathrm{Bias}(\hat{Y}_i^{\mathrm{IAUM}}|\mathbf{X}_i, W_i=1)^2 - \mathrm{Bias}(\hat{Y}_i^{\mathrm{DR}}|\mathbf{X}_i, W_i=1)^2 \\
=& (\delta_i^0 + \big(1-\hat{e}(\mathbf{X}_i)\big)\delta_i^1)^2 - (\delta_i^0 + \frac{1-\hat{e}(\mathbf{X}_i)}{\hat{e}(\mathbf{X}_i)}\delta_i^1)^2 \\
=& -2\frac{(1-\hat{e}(\mathbf{X}_i))^2}{\hat{e}(\mathbf{X}_i)}\delta_i^1 * (\delta_i^0 + \frac{1-\hat{e}(\mathbf{X}_i)^2}{2\hat{e}(\mathbf{X}_i)}\delta_i^1) \\
=& -2\kappa_0\delta_i^1 * (\delta_i^0 + \kappa_1\delta_i^1) \\
=& -2\kappa_0(\delta_i^1)^2 * (\kappa_1 + \frac{\delta_i^0}{\delta_i^1}) \\
\leq& -2\kappa_0(\delta_i^1)^2 * (\kappa_1 - |\frac{\delta_i^0}{\delta_i^1}|) < 0,
\end{aligned} \tag{17}$$

where $\kappa_0 = \frac{(1-\hat{e}(\mathbf{X}_i))^2}{\hat{e}(\mathbf{X}_i)} > 0$ and $\kappa_1 = \frac{1-\hat{e}(\mathbf{X}_i)^2}{2\hat{e}(\mathbf{X}_i)}$. Given $|\frac{\delta_i^0}{\delta_i^1}| < 1$, we can get $\kappa_1 > 0$ when $\hat{e}(\mathbf{X}_i) < \sqrt{2} - 1 \approx 0.414$, which is obviously satisfied on an imbalanced dataset. Therefore, we can get $\triangle < 0$ and prove $\mathrm{Bias}(\hat{Y}_i^{\mathrm{IAUM}}|\mathbf{X}_i, W_i=1) < \mathrm{Bias}(\hat{Y}_i^{\mathrm{DR}}|\mathbf{X}_i, W_i=1)$, and the case when $W_i = 0$ can be derived in a similar way.

As proved in (Saito, Sakata, and Nakata 2019), DR would has smaller bias than TOM when $|\Delta_i^{(k)}| < \mu_i^{(k)} (\forall k \in \{0,1\})$, which is a reasonable condition to be satisfied due to the powerful fitting ability of existing machine learning algorithms. Therefore, we can finally get $\mathrm{Bias}(\hat{Y}_i^{\mathrm{IAUM}}|\mathbf{X}_i, W_i) < \mathrm{Bias}(\hat{Y}_i^{\mathrm{DR}}|\mathbf{X}_i, W_i) < \mathrm{Bias}(\hat{Y}_i^{\mathrm{TOM}}|\mathbf{X}_i, W_i)$. □

Deviation Analysis of IAUM

In this subsection, we analyze the deviation of IAUM, and then compare it with existing works.

Theorem 4. *Given the propensity score $e(\mathbf{X}_i)$, with probability $1-\eta$, the following inequation holds:*

$$\begin{aligned}
&|\mathbf{G}^* - \mathbb{E}[\mathbf{G}^*]| \\
\leq& \sqrt{\frac{1}{2N^2}\log(\frac{2}{\eta})\sum_{i=1}^{n}(\hat{Y}_i^*(W_i=1) - \hat{Y}_i^*(W_i=0))^2},
\end{aligned} \tag{18}$$

where $\hat{Y}_i^*$ is the proxy/transformed outcome generated by any proxy method, and $*$ can be TOM, DR or IAUM. $\hat{Y}_i^*(W_i=1)$ and $\hat{Y}_i^*(W_i=0)$ denote the proxy outcome when $W_i=1$ and $W_i=0$, separately. $\mathbf{G}^* = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i^* - \tau_i)$.

Proof. Since we assume that each observation indicator W follows the Bernoulli distribution with probability $e(x)$ (that is the propensity score), we can rewrite $\hat{Y}_i^*$ as follows:

$$\begin{cases} P(\hat{Y}_i^*(W_i=1)|\mathbf{X}_i) = e_i \\ P(\hat{Y}_i^*(W_i=0)|\mathbf{X}_i) = 1-e_i. \end{cases} \tag{19}$$

Random variable $\hat{Y}_i^* - \tau_i$ takes the value either $\hat{Y}_i^*(W_i=0) - \tau_i$ or $\hat{Y}_i^*(W_i=1) - \tau_i$, which still follow the Bernoulli distribution.

$$\begin{cases} P(\hat{Y}_i^*(W_i=1) - \tau_i|\mathbf{X}_i) = e_i \\ P(\hat{Y}_i^*(W_i=0) - \tau_i|\mathbf{X}_i) = 1-e_i. \end{cases} \tag{20}$$

Therefore, according to Hoeffding's inequality (Hoeffding 1994), with probability $1-\eta$, for any $\hat{\xi} > 0$ we have the following inequality:

$$P(|\sum_i [\hat{Y}_i^* - \tau_i] - \mathbb{E}[\sum_i \hat{Y}_i^* - \tau_i]| \geq \hat{\xi}) \leq 2\exp(\frac{-2\hat{\xi}^2}{\sum_i \rho^2}), \tag{21}$$

where ρ is equal to $\hat{Y}_i^*(W_i=1) - \tau_i - (\hat{Y}_i^*(W_i=0) - \tau_i) = \hat{Y}_i^*(W_i=1) - \hat{Y}_i^*(W_i=0)$.

The summation here is to sum over all samples $i \in N$. We set $\hat{\xi} = \xi|N|(\xi > 0 \Leftrightarrow \hat{\xi} > 0)$. Based on the above inequality, we can get the inequality of the $\mathbf{G}^*$ as follows:

$$P(|\mathbf{G}^* - \mathbb{E}[\mathbf{G}^*]| \geq \xi) \leq 2\exp(\frac{-2\xi^2|N|^2}{\sum_i \rho_i^2}). \tag{22}$$

Setting the right side of the inequality to the probability η and solving for ξ to complete the proof of Equation (18). □

To simplify the formula, we can set a substitute variable C as $\frac{1}{2N^2}\log(\frac{2}{\eta})$, and let Λ^* denotes the upper bound of the deviation. With the above theorem, we can derive the Λ^* of our proposed IAUM method as:

$$\begin{aligned}
&\Lambda^{\mathrm{IAUM}} \\
=& \sqrt{C\sum_{i=1}^{n}\big((2-\hat{e}(\mathbf{X}_i))(Y_i^1 - \hat{\mu}_i^1) + (1+\hat{e}(\mathbf{X}_i)(Y_i^0 - \hat{\mu}_i^0)\big)^2} \\
=& \sqrt{C\sum_{i=1}^{n}\big((2-\hat{e}(\mathbf{X}_i))\Delta_i^1 + (1+\hat{e}(\mathbf{X}_i)\Delta_i^0\big)^2}.
\end{aligned} \tag{23}$$

Similarly, given Theorem 4 and the definition of TOM and DR estimator, we can derive the deviation of TOM (Equation (10)) and DR (Equation (12)). We can easily prove that $2-\hat{e}(\mathbf{X}_i) < \frac{1}{\hat{e}(\mathbf{X}_i)}$ and $1+\hat{e}(\mathbf{X}_i) < \frac{1}{1-\hat{e}(\mathbf{X}_i)}$, therefore $\Lambda^{\mathrm{IAUM}} < \Lambda^{\mathrm{DR}} < \Lambda^{\mathrm{TOM}}$.

Experiment

In this section, we conduct the experiments on synthetic and industrial datasets to validate the following. (1) Compared with other propensity score based methods, our IAUM method has the smallest bias and variance in uplift estimation. (2) In the highly imbalanced dataset, the prediction difficulties of treatment/control outcome vary a lot. (3) Our proposed method can efficiently target the audience and obtain the highest return of investment on the industrial application.

Datasets

We conduct the experiment on three datasets: (1) Synthetic dataset; (2) Industrial dataset; (3) Right Heart Catheterization (RHC) dataset.

Synthetic Dataset. First, we evaluated the performance of our method and other existing method on synthetic datasets composed of eight scenarios. Each scenario was defined by the data generating processes used in (Saito, Sakata, and Nakata 2019; Schuler et al. 2018). Specifically, we generate data composed of one million individuals with 10 features, which follow a Gaussian distribution with a mean of 0 and a variance of 1. Then we set the probability of the individual to receive treatment around 0.1 to construct the imbalance of the dataset. The characteristics of the eight scenarios are briefly summarized in Table 1. $\mathbb{E}[Y_i^1]$ and $\mathbb{E}[Y_i^0]$ represent the mean of τ_i. In each scenario, the data is split into training set and test set with the ratio of $50\%/50\%$.

Industrial Dataset. To further evaluate the effectiveness of the proposed method, we compare these methods on an industrial dataset collecting from a real mobile marketing campaign. Here the treatment is to expose an advertisement to the logged user for promoting conversion, and the observed outcome is whether the user converts within this login. $\mathbf{X}$ in this dataset is the feature that encodes the information of users' demographic profiles and online behaviors. One thing to note is that the user can achieve the conversion via other approaches except for the advertisement, i.e., the control group can also be observed positive outcome. Therefore, through modeling the uplift of the treatment, we can target at the users with large uplift to save the budget on the advertising. Here we use 7-day data that includes millions of users to construct the dataset, where the first 6 days are for training, and the data collected on the last day is for testing. In addition, the propensity score $e(\mathbf{X}_i)$ is smaller than 5%, which indicates that this dataset is highly imbalanced.

RHC Dataset. We chose Right Heart Catheterization (RHC) data (Saito, Sakata, and Nakata 2019) as the real-world data set to compare our procedure with existing methods. RHC is the diagnosis of critically ill patients, and the data set contains 5735 patients. In this dataset, 2184 patients received treatment and 3551 did not receive treatment, and this treatment allocation is not random. In order to test all methods under an imbalance setting, we build an imbalanced RHC dataset via randomly down-sampling the treatment group to $e(\mathbf{X}) = 0.1$.

Comparison Methods

We compare our method with several baselines:

- **TMA** (Jaskowski and Jaroszewicz 2012): an estimator defines ITE as the difference between predicted outcomes coming from two group of subjects;
- **TOM** (Athey and Imbens 2015): an estimator based on the transformed outcome via reweighting based on inverse propensity score;
- **DR** (Funk et al. 2011): a doubly robust estimator that combines error imputation based estimator and inverse propensity score estimator;
- **X-Learner** (Künzel et al. 2019): a two model approach that crossovers the information in the treated and control subjects;
- **SDRM** (Saito, Sakata, and Nakata 2019): an estimator that switches between doubly robust estimator and predicted $\hat{\mu}_i^1 - \hat{\mu}_i^0$;
- **TMLE** (Schuler and Rose 2017): a targeted maximum likelihood estimator;
- **TDVAE** (Zhang, Liu, and Li 2020): a variational inference approach to simultaneously infer latent factors from the observed variables.

Model Setup

On the synthetic dataset and the RHC dataset, we use the linear regression as the base learners for simplicity. For each scenario, we repeat the training process ten times and report the average bias and variance of the deviation between the expected true value and the predictions of the model output.

As the industrial dataset has high-dimensional features, we choose the multilayer perceptron (MLP) with three hidden layers (the number of neurons is 512, 128 and 128, respectively) as the base learner to fit the data. All neural network-based methods are optimized by Adam (Kingma and Ba 2014) optimizer with a learning rate of $3e-4$, and set the batch size to 512.

Evaluation Metrics

Synthetic Dataset. On the synthetic dataset, since potential outcomes are simulated based on covariates in a carefully designed way, the ground truth ITEs are known. Therefore, we can directly calculate the bias and variance of the deviation between the predictions and the ground truth. The smaller the bias and variance is, the better the performance is.

Industrial & RHC Dataset. On the industrial dataset and RHC dataset, only one of the potential outcomes is observable, and the ground truth ITEs are not available. Therefore, we adopt the widely used metrics, the Qini curve (Radcliffe 2007) and the area under the Qini curve (Qini coefficient), to evaluate the performance of different estimators. Specifically, the Qini curve is defined as:

$$\text{Qini}(\phi) = \frac{N_{y=1}^{w=1}(\phi)}{N_t} - \frac{N_{y=1}^{w=0}(\phi)}{N_c}, \tag{24}$$

where ϕ is the fraction of population treated ordered by predicted uplift (from highest to lowest). $N_{y=1}^{w=1}(\phi)$ and $N_{y=1}^{w=0}(\phi)$ are the count of positive outcomes in the treatment and control groups respectively from ϕ. N_t and N_c are the numbers of subjects within the entire treatment and control groups, respectively. The larger the Qini coefficient is, the better the performance is.

Results on the Synthetic Dataset

Figure 1 shows the bias of our proposed methods as well as the methods adopting the propensity score to construct proxy outcome. It is observed that among all datasets, the proposed IAUM has the smallest bias. And TOM shows

No.	1	2	3	4	5	6	7	8
$\mathbb{E}[Y_i^1]$	5.998	-1.931	6.008	0.995	-4.999	0.312	-1.349	3.617
$\mathbb{E}[Y_i^0]$	6.002	-2.075	9.998	-7.002	-3.000	-0.317	1.660	-2.362
Mean of τ_i	0.000	0.159	-3.999	7.998	-1.997	0.635	-3.001	6.001

Table 1: Characteristics of eight scenarios.

large bias on these datasets, which suggests that TOM is prone to be biased is due to the biased estimated propensity score. With the technique of switching, SDRM successfully reduced the bias, but its bias is still larger than IAUM on the imbalanced setting. Similar trend can be observed in terms variance, as shown in Figure 2. Overall, the proposed method IAUM consistently outperforms other methods on the synthetic dataset with the smallest bias and variance.

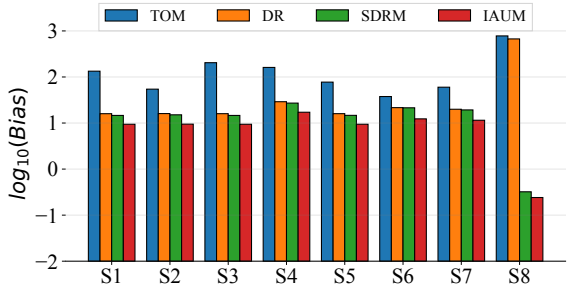

Figure 1: The log-scaled bias on the synthetic dataset.

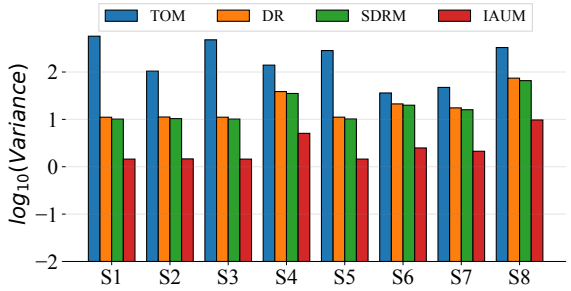

Figure 2: The log-scaled variance on the synthetic dataset.

To further demonstrate the superiority of our proposed IAUM method, we vary the ratio of treated group in the dataset from 0.1 to 0.5. The ratio can be viewed as the imbalance level of the dataset, and the farther the ratio away from 0.5, the higher level of the imbalance. Figure 3 show the results of IAUM and other baselines over different group size ratio. Due to the space limit, we only report the results on scenario 8, and similar trends can be observed in other scenarios. It is observed that all methods have similar performance when the group size is around 0.5. With the increase of the imbalance level, IAUM consistently outperforms other methods with a significant gap, which reflects the effectiveness of our proposed method.

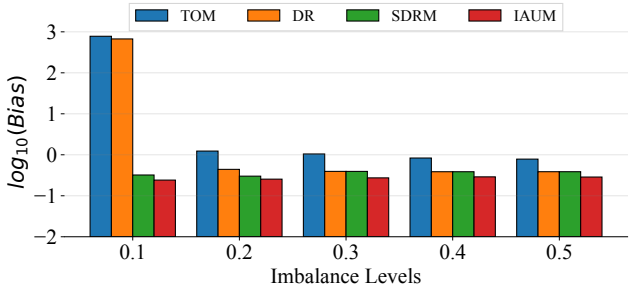

Figure 3: The log-scaled bias of each method under different data imbalance levels.

Additionally, to show the rationality of our proxy outcome design, in Figure 4, we report the mean absolute error of the potential outcome model MAE(μ_i^1) and MAE(μ_i^0) under different imbalance levels ranging from 0.1 to 0.5. It can be seen from the figure that the MAE of the μ_0 model is significantly smaller than the MAE of the μ_1 model, indicating that predicting the treated outcome is much more difficult than the control outcome when the treated group size is extremely small. This observation validates that it is reasonable to use the propensity score as the weight, which prevents the constructed proxy outcome from assigning high weights to the part containing $\hat{\mu}_1$.

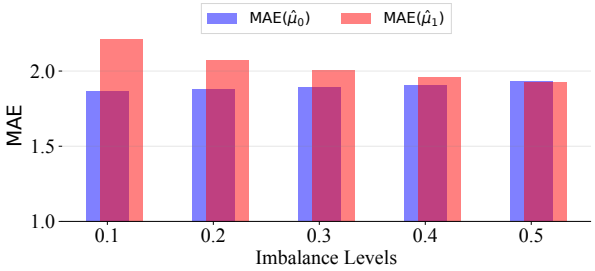

Figure 4: The outcome prediction results under different data imbalance levels.

TMA	TOM	DR	X-Learner	SDRM	IAUM
0.076	0.095	0.149	0.110	0.468	**0.593**

Table 2: Qini coefficients on the industrial dataset.

Results on the Industrial and RHC Dataset

Figure 5 shows Qini curves on the industrial dataset. We can see that IAUM outperforms other uplift modeling methods with a large margin, and SDRM is relatively closed to IAUM due to its powerful switching technique. TOM, X-Learner, TMA and DR show comparable performance, which are both better than the randomized estimation. Moreover, Table 2 presents the results of Qini coefficients, and IAUM has the largest Qini coefficient among these methods, which well verifies its effectiveness. According to Figure 5, we can consider users with the top-scored 20% uplift as the target audience to maximize the return of investment since most of accumulated gain is obtained on the top 20% of treated individuals.

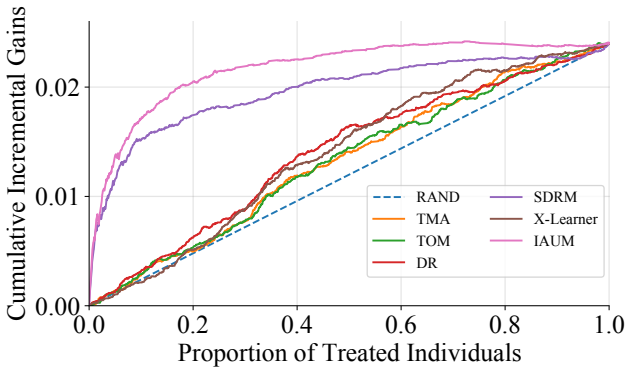

Figure 5: Qini curves on the industrial dataset.

TMA	TOM	DR	X-Learner	SDRM	TMLE	TDVAE	IAUM
0.0059	0.0078	0.0116	0.0114	0.0123	0.0095	0.0130	**0.0143**

Table 3: Qini coefficients on the imbalanced RHC dataset.

Table 3 presents the results of Qini coefficients on the RHC dataset, and we can see a similar performance with the industrial dataset. TMA and TOM both perform poor under the imbalance setting, and the other baseline methods with comparable performance still fail to outperform IAUM, which has the largest Qini coefficient.

Conclusion

In this paper, through theoretical and quantitative analysis, we prove that existing uplift modeling methods would suffer from large bias and deviation on a highly imbalanced dataset. To overcome this drawback, we propose an imbalance-aware uplift modeling method via constructing a robust proxy outcome, which adaptively combines the doubly robust estimator and the imputed treatment effects based on the propensity score. Experimental results well demonstrated the effectiveness of the proposed method, and show its power in the industrial setting. Future work will focus on how to construct robust proxy outcome while considering the deviation of the propensity score since poor estimation of the propensity score would lead to large errors.

Appendix

Proof of Theorem 1

Proof. Given the definition of TOM estimator (Equation (7) in our paper), the bias of TOM is:

$$\begin{aligned}
\mathrm{Bias}(\hat{Y}_i^{\mathrm{TOM}}|X_i, W_i=1) &= |\mathbb{E}[\hat{Y}_i^{\mathrm{TOM}}|X_i, W_i=1] - \tau_i| \\
&= \big|\mathbb{E}[\frac{Y_i^1}{\hat{e}(X_i)}|X_i] - (\mu_i^1 - \mu_i^0)\big| \\
&= \big|\frac{\mu_i^1}{\hat{e}(X_i)} - (\mu_i^1 - \mu_i^0)\big| \\
&= \big|\mu_i^0 + \frac{1-\hat{e}(X_i)}{\hat{e}(X_i)}\mu_i^1\big|. \\
\mathrm{Bias}(\hat{Y}_i^{\mathrm{TOM}}|X_i, W_i=0) &= |\mathbb{E}[\hat{Y}_i^{\mathrm{TOM}}|X_i, W_i=0] - \tau_i| \\
&= \big|\mathbb{E}[\frac{-Y_i^0}{1-\hat{e}(X_i)}|X_i] - (\mu_i^1 - \mu_i^0)\big| \\
&= \big|\frac{-\mu_i^0}{1-\hat{e}(X_i)} - (\mu_i^1 - \mu_i^0)\big| \\
&= \big|\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}\mu_i^0 + \mu_i^1\big|.
\end{aligned} \tag{25}$$

□

Proof of Theorem 2

Proof. Given the definition of DR estimator (Equation (8) in our paper), the bias of DR is:

$$\begin{aligned}
&\mathrm{Bias}(\hat{Y}_i^{\mathrm{DR}}|X_i, W_i=1) \\
=&|\mathbb{E}[\hat{Y}_i^{\mathrm{DR}}|X_i, W_i=1] - \tau_i| \\
=&\big|\mathbb{E}[\hat{\mu}_i^1 - \hat{\mu}_i^0 + \frac{Y_i^1 - \hat{\mu}_i^1}{\hat{e}(X_i)}|X_i] - (\mu_i^1 - \mu_i^0)\big| \\
=&\big|\hat{\mu}_i^1 - \hat{\mu}_i^0 + \frac{\mu_i^1 - \hat{\mu}_i^1}{\hat{e}(X_i)} - (\mu_i^1 - \mu_i^0)\big| \\
=&\big|(\hat{\mu}_i^0 - \mu_i^0) + \frac{1-\hat{e}(X_i)}{\hat{e}(X_i)}(\hat{\mu}_i^1 - \mu_i^1)\big| \\
=&\big|\delta_i^0 + \frac{1-\hat{e}(X_i)}{\hat{e}(X_i)}\delta_i^1\big|.
\end{aligned} \tag{26}$$

$$\begin{aligned}
&\mathrm{Bias}(\hat{Y}_i^{\mathrm{DR}}|X_i, W_i=0) \\
=&|\mathbb{E}[\hat{Y}_i^{\mathrm{DR}}|X_i, W_i=0] - \tau_i| \\
=&\big|\mathbb{E}[\hat{\mu}_i^1 - \hat{\mu}_i^0 - \frac{Y_i^0 - \hat{\mu}_i^0}{1-\hat{e}(X_i)}|X_i] - (\mu_i^1 - \mu_i^0)\big| \\
=&\big|\hat{\mu}_i^1 - \hat{\mu}_i^0 - \frac{\mu_i^0 - \hat{\mu}_i^0}{1-\hat{e}(X_i)} - (\mu_i^1 - \mu_i^0)\big| \\
=&\big|\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}(\hat{\mu}_i^0 - \mu_i^0) + (\hat{\mu}_i^1 - \mu_i^1)\big| \\
=&\big|\frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}\delta_i^0 + \delta_i^1\big|.
\end{aligned}$$

□

References

Athey, S.; and Imbens, G. W. 2015. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5): 1–26.

Chen, X.; Liu, Z.; Yu, L.; Li, S.; Gu, L.; Zeng, X.; Tan, Y.; and Gu, J. 2021. Adversarial Learning for Incentive Optimization in Mobile Payment Marketing. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2940–2944.

Funk, M. J.; Westreich, D.; Wiesen, C.; Stürmer, T.; Brookhart, M. A.; and Davidian, M. 2011. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7): 761–767.

Gutierrez, P.; and Gérardy, J.-Y. 2017. Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications and APIs*, 1–13. PMLR.

Hansotia, B.; and Rukstales, B. 2002. Incremental value modeling. *Journal of Interactive Marketing*, 16(3): 35.

Hoeffding, W. 1994. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, 409–426. Springer.

Imai, K.; and Ratkovic, M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1): 443–470.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jaskowski, M.; and Jaroszewicz, S. 2012. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, volume 46.

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

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

Künzel, S. R.; Stadie, B. C.; Vemuri, N.; Ramakrishnan, V.; Sekhon, J. S.; and Abbeel, P. 2018. Transfer learning for estimating causal effects using neural networks. *arXiv preprint arXiv:1808.07804*.

Li, S.; Lv, F.; Jin, T.; Lin, G.; Yang, K.; Zeng, X.; Wu, X.-M.; and Ma, Q. 2021. Embedding-Based Product Retrieval in Taobao Search. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3181—-3189.

Lo, V. S. 2002. The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, 4(2): 78–86.

Ma, Q.; Li, S.; and Cottrell, G. 2020. Adversarial Joint-Learning Recurrent Neural Network for Incomplete Time Series Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Nichols, A. 2007. Causal inference with observational data. *The Stata Journal*, 7(4): 507–541.

Olaya, D.; Coussement, K.; and Verbeke, W. 2020. A survey and benchmarking study of multitreatment uplift modeling. *Data Mining and Knowledge Discovery*, 34(2): 273–308.

Radcliffe, N. 2007. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, 14–21.

Saito, Y.; Sakata, H.; and Nakata, K. 2019. Doubly robust prediction and evaluation methods improve uplift modeling for observational data. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, 468–476. SIAM.

Schuler, A.; Baiocchi, M.; Tibshirani, R.; and Shah, N. 2018. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*.

Schuler, M. S.; and Rose, S. 2017. Targeted maximum likelihood estimation for causal inference in observational studies. *American journal of epidemiology*, 185(1): 65–73.

Wang, X.; Zhang, R.; Sun, Y.; and Qi, J. 2019. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, 6638–6647. PMLR.

Yao, L.; Chu, Z.; Li, S.; Li, Y.; Gao, J.; and Zhang, A. 2020. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, 15(5): 46.

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.

Yao, L.; Li, S.; Li, Y.; Huai, M.; Gao, J.; and Zhang, A. 2019. Ace: Adaptively similarity-preserved representation learning for individual treatment effect estimation. In *2019 IEEE International Conference on Data Mining (ICDM)*, 1432–1437. IEEE.

Yao, L.; Li, Y.; Li, S.; Huai, M.; Gao, J.; and Zhang, A. 2021. SCI: Subspace Learning Based Counterfactual Inference for Individual Treatment Effect Estimation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3583–3587. ACM.

Yu, L.; Wu, Z.; Cai, T.; Liu, Z.; Zhang, Z.; Gu, L.; Zeng, X.; and Gu, J. 2021. Joint Incentive Optimization of Customer and Merchant in Mobile Payment Marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15000–15007.

Zaniewicz, Ł.; and Jaroszewicz, S. 2013. Support vector machines for uplift modeling. In *2013 IEEE 13th International Conference on Data Mining Workshops*, 131–138. IEEE.

Zhang, W.; Le, T. D.; Liu, L.; Zhou, Z.-H.; and Li, J. 2017. Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*, 33(15): 2372–2378.

Zhang, W.; Li, J.; and Liu, L. 2020. A unified survey on treatment effect heterogeneity modeling and uplift modeling. *arXiv preprint arXiv:2007.12769*.

Zhang, W.; Liu, L.; and Li, J. 2020. Treatment effect estimation with disentangled latent factors. *arXiv preprint arXiv:2001.10652*.