Explicit Feature Interaction-aware Uplift Network for Online Marketing

Dugang Liu*
Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University
Shenzhen, China
dugang.ldg@gmail.com

Xing Tang†
FiT, Tencent
Shenzhen, China
shawntang@tencent.com

Han Gao
FiT, Tencent
Shenzhen, China
hansologao@tencent.com

Fuyuan Lyu
McGill University
Montreal, Canada
fuyuan.lyu@mail.mcgill.ca

Xiuqiang He†
FiT, Tencent
Shenzhen, China
xiuqianghe@tencent.com

ABSTRACT

As a key component in online marketing, uplift modeling aims to accurately capture the degree to which different treatments motivate different users, such as coupons or discounts, also known as the estimation of individual treatment effect (ITE). In an actual business scenario, the options for treatment may be numerous and complex, and there may be correlations between different treatments. In addition, each marketing instance may also have rich user and contextual features. However, existing methods still fall short in both fully exploiting treatment information and mining features that are sensitive to a particular treatment. In this paper, we propose an explicit feature interaction-aware uplift network (EFIN) to address these two problems. Our EFIN includes four customized modules: 1) a feature encoding module encodes not only the user and contextual features, but also the treatment features; 2) a self-interaction module aims to accurately model the user's natural response with all but the treatment features; 3) a treatment-aware interaction module accurately models the degree to which a particular treatment motivates a user through interactions between the treatment features and other features, i.e., ITE; and 4) an intervention constraint module is used to balance the ITE distribution of users between the control and treatment groups so that the model would still achieve a accurate uplift ranking on data collected from a non-random intervention marketing scenario. We conduct extensive experiments on two public datasets and one product dataset to verify the effectiveness of our EFIN. In addition, our EFIN has been deployed in a credit card bill payment scenario of a large online financial platform with a significant improvement.

*This work was done during his internship at FiT, Tencent.
†Co-corresponding authors

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599820

CCS CONCEPTS

• **Information systems** → **Personalization**; • **Applied computing** → **Economics**.

KEYWORDS

Uplift modeling, Feature interaction, Treatment-aware interaction, Intervention constraint

ACM Reference Format:
Dugang Liu, Xing Tang, Han Gao, Fuyuan Lyu, and Xiuqiang He. 2023. Explicit Feature Interaction-aware Uplift Network for Online Marketing. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3580305.3599820

1 INTRODUCTION

To increase the user engagement and platform revenue, providing some specific incentives to the users, such as coupons [37], discounts [19], and bonuses [1], is an important strategy in online marketing [27]. Since these incentives usually have a cost and different users have different responses to these incentives, such as some users do not consume without a coupon and some users will consume anyway, how to accurately identify the corresponding sensitive user groups for each incentive is critical to maximize marketing benefits [14, 34]. To achieve this goal, we need to accurately capture the difference between users' responses to various incentives compared to those without incentives. Unlike traditional supervised learning, this involves a typical causal inference problem, because in a practical scenario, we can usually only observe one type of the user responses, which may be for a certain incentive (i.e., treatment group) or for no incentive (i.e., control group). Therefore, the change in the user's response caused by different incentives (or treatments) that we want to obtain can be regarded as the estimation of the individual treatment effect (ITE) [36], also known as the uplift. To solve the above estimation problem, in recent years, uplift modeling has been proposed and its effectiveness has been verified [5, 7, 10].

The existing uplift modeling methods mainly includes three research lines according to the design ideas: 1) *Meta-learner based.* The basic idea of this line is to use the existing prediction methods

to build the estimator for the users' responses, which may be global (i.e., S-Learner) or divided by the treatment and control groups (i.e., T-Learner) [17]. Based on this, different two-step learners can be designed by introducing various additional operations, such as X-Learner [17], R-Learner [24], and DR-Learner [4], etc. 2) *Tree based.* The basic idea of this line is to use a tree structure to gradually divide the entire user population into the sub-populations that are sensitive to each treatment. The key step is to directly model the uplift using different splitting criteria, such as based on various distribution divergences [25] and the expected responses [29, 38]. In addition, causal forest [3] obtained by integrating multiple trees is another representative method on this line, and several variants have been proposed [1, 32]. 3) *Neural network based.* The basic idea of this line is to take advantage of neural networks to design more complex and flexible estimators for the user's response [16, 21, 35, 39], and most of them can be seen as improvements of the T-learner [8, 9, 30, 31]. In this paper, we focus on neural network-based line because it can be better adapted to the goal of feature interaction modeling introduced in this paper due to the flexibility of neural networks. Also, since various neural network models are commonly employed in commercial systems, research on this line can be more easily integrated than other lines. We present the architectures of some representative methods in neural network-based uplift modeling in Figure 1.

Although existing uplift modeling methods have shown promising results, most of them still fall short in both fully exploiting treatment information and mining features that are sensitive to a particular treatment. In an online marketing, the treatment usually has many features that describe it in detail in addition to the index ID. For example, a coupon may include a specific amount and a minimum spending amount to be reached. This also means that different treatments may be related, such as having similar amounts or minimum spending to be achieved. Intuitively, this information is beneficial for obtaining a accurate uplift, e.g., the correlation between the treatments can prompt the model to discover that a user's response to a coupon worth 1000 should be more similar to a coupon worth 900 than to a coupon worth 100. However, as shown in Figure 1, we can find that almost all related methods do not explicitly utilize treatment features, which may be detrimental to the uplift estimation. We refer to this challenge as **underutilization of treatment features**. Furthermore, the above challenges will also prevent most related methods from accurately capturing the sensitive features associated with each treatment, due to the lack of modeling of the interactions between treatment features and the rest. We refer to this challenge as **underutilization of feature interactions**. Note that explicitly modeling the treatment features may also make the model compatible with a variety of marketing scenarios, where treatment options may be binary, multi-valued, or continuous, without significantly increasing the size of the model.

To address the above two challenges, in this paper, we propose an explicit feature interaction-aware uplift network (EFIN). Specifically, our EFIN includes four modules: 1) a feature encoder module aims to encode a marketing instance containing the user features, the contextual features, and the treatment features; 2) a self-interaction module is responsible for the responses of the users in the control group. It uses a self-attention network to model the interactions between all the features except the treatment features to capture a subset of features associated with the natural responses (i.e., not receiving any the treatment); 3) a treatment-aware interaction module is responsible for the responses of users in the treatment group. It uses a treatment-aware attention network to model the interaction between the treatment features and other features to identify subsets of features that are sensitive to different treatments, and to accurately capture a user's changes in response to different treatments; and 4) an intervention constraint module is used to balance the ITE distribution of users between the treatment and control groups so that our EFIN could be more robust in different scenarios. This module is necessary since the treatment assignment is usually non-random in a real marketing scenario and will result in differences in user distribution between control and treatment groups. Finally, we conduct extensive offline and online evaluations and the results validate the effectiveness of our EFIN.

2 RELATED WORK

In this section, we briefly review some relevant works on two research topics, including uplift modeling and feature interaction.

2.1 Uplift Modeling

Uplift modeling aims to identify the corresponding sensitive population for each specific treatment by accurately estimating ITE. The existing uplift modeling methods mainly includes three research lines: 1) a meta-learner-based method focuses on using existing prediction methods to learn a one-step learner [17] or a two-step learner [4, 24] for the user's response, where the treatment information is usually integrated as one-dimensional discrete features or as a prior for switching prediction branches. 2) a tree-based method employs a specific tree or a forest structure with splitting criterion of different metrics to gradually divide the sensitive subpopulations corresponding to each treatment from the entire population [3, 25, 38], where the treatment information is included in the calculation of the splitting process; and 3) a neural network-based method combines the advantages of neural networks to introduce some more complex and flexible architectures to model the response process to the treatment, which can learn a more accurate estimator for the users' responses or the uplifts. Furthermore, there are only a few works that address uplift modeling by linking it to the well-established problems in other fields, such as the knapsack problem [2, 12]. Our EFIN follows a neural network-based line, but differs significantly from existing related works, especially in the explicit utilization of the treatment feature and the modeling of its interactions with other features.

2.2 Feature Interaction

Feature interactions are designed to model combinations between different features and have been shown to significantly improve the performance of a response model [22, 23]. Existing feature interaction methods can be mainly divided into three categories, including second-order interactions, higher-order interactions, and structural interactions. In second-order interactions, the inner product between the embedding representations of two features is usually

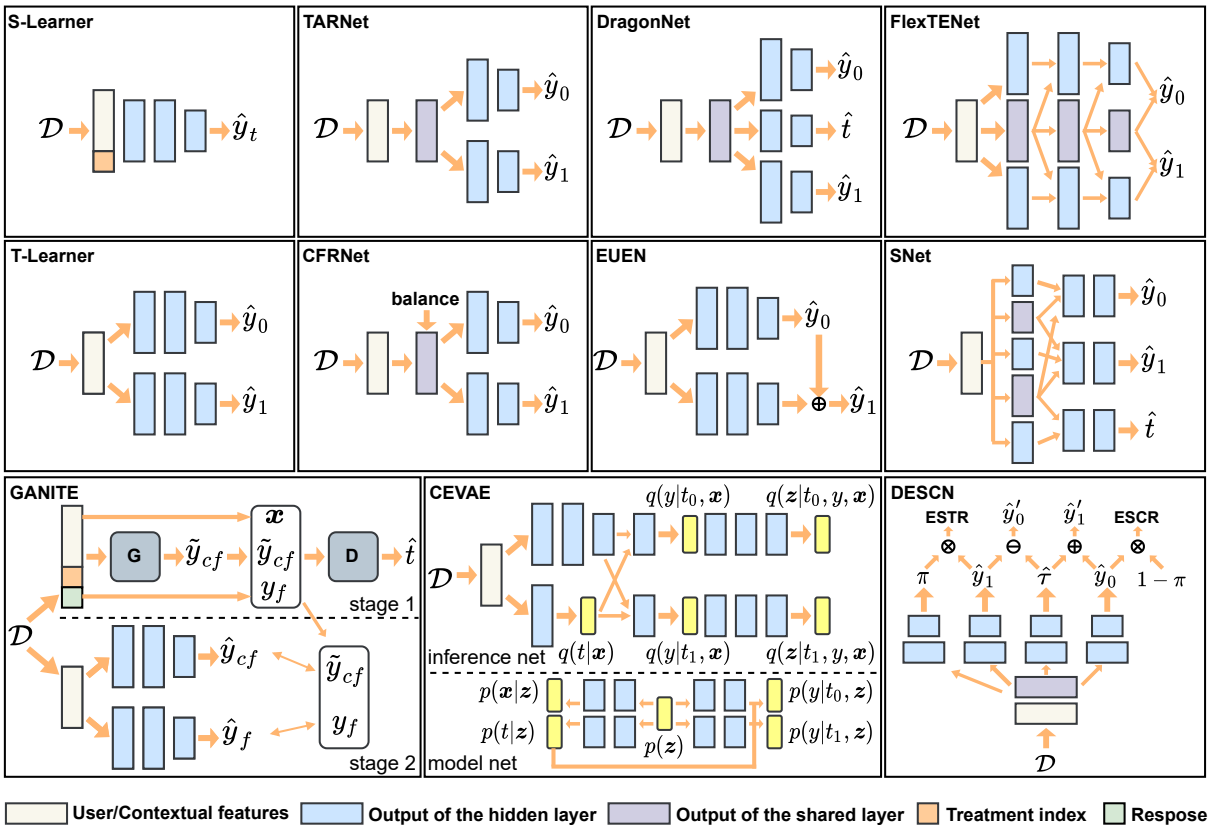

Figure 1: Architecture diagram of some representative methods in neural network-based uplift modeling, where $\mathcal{D}$ denotes a training set, $\hat{y}_0$ and $\hat{y}_1$ denote the predicted response for the control and treatment groups, respectively, and $\hat{t}$ denotes the predicted label of the treatment. In GANITE, G, D, $\widetilde{y}_{cf}$, and $\hat{y}_{cf}$ are the generator, the discriminator, the generated counterfactual response and the corresponding predicted response, respectively. In CEVAE, $p()$ and $q()$ denote different distributions in the inference network and the model network, respectively. In DESCN, $\hat{\tau}$, π, $\hat{y}'_0$, and $\hat{y}'_1$ denote the predicted ITE, the probability that an instance belongs to the treatment group, and the predicted cross-control and cross-treatment responses, respectively.

considered, and factorization machines and their variants are representative methods [15, 26]. Modeling of higher-order interactions relies on neural networks, and many architectures have been proposed to enhance model performance, interpretability, and efficient fusion of lower- and higher-order interactions [13, 33]. In addition, based on the graph structure, some methods aim to exploit the additional structural information to further improve higher-order interactions [18, 20]. Although feature interaction has achieved success on many tasks, research on its application in uplift modeling is still lacking. Our EFIN aims to bridge the gap in this research direction.

3 PRELIMINARIES

Let $\{z_i\}_{i=1}^n = \{(\boldsymbol{x}_i, \boldsymbol{t}_i, y_i)\}_{i=1}^n \in \mathcal{X} \times \mathcal{T} \times \mathcal{Y}$ denote a marketing instance, where $\boldsymbol{x}_i = \left[x_{i0}, x_{i1}, \ldots, x_{id_{x-1}}, x_{id_x}\right]$ is the d_x-dimensional user features and contextual features included in the i-th instance, $\boldsymbol{t}_i = \left[t_{i0}, t_{i1}, \ldots, t_{id_{t-1}}, t_{id_t}\right]$ is the d_t-dimensional treatment features included, n is the number of the training instances, and $y_i \in \{0, 1\}$ is the response label for the i-th instance. Without loss of generality, we assume that the first treatment feature of each instance denotes the index ID of the treatment, and the total number of treatments is K, i.e., $t_{i0} \in \{0, 1, \ldots, K\}$.

Following the Neyman-Rubin potential outcome framework [28], let $y_i(k)$ and $y_i(0)$ denote the potential outcome when the user in the i-th instance gets a particular treatment $k \in \{1, \ldots, K\}$ or is not treated, respectively. The probability of each treatment being assigned can be denoted as $\pi_k(\boldsymbol{x}_i) = P(t_{i0} = k|\boldsymbol{x}_i)$, also known as a propensity score. Since we can usually only observe $y_i(k)$ or $y_i(0)$, but not both, i.e., $y_i = y_i(k)$ or $y_i = y_i(0)$, there is no true uplift result $y_i(k) - y_i(0)$ for each instance, which is a key reason uplift modeling differs from traditional supervised learning. Therefore, uplift modeling aims to accurately estimate the expected individual treatment effect $\tau_k(\boldsymbol{x}_i)$ for each instance. Specifically, following standard assumptions [39], this estimate can be expressed as,

$$\begin{aligned}\tau_k(x_i) &= \mathbb{E}(y_i(k) - y_i(0)|x_i), \\ &= \mathbb{E}(y_i(k)|t_{i0} = k, x_i) - \mathbb{E}(y_i(0)|t_{i0} = 0, x_i).\end{aligned} \tag{1}$$

After obtaining all the estimated individual treatment effects $\tau_k(x_i)$, we can rank them and make a rational treatment assignment.

4 THE PROPOSED METHOD

4.1 Architecture

As mentioned in Section 1, most of the existing methods generally suffer from two challenges of underutilization of treatment features and feature interactions. To address the above two challenges, in this paper, we propose an explicit feature interaction-aware uplift network (EFIN) and illustrate the architecture of our EFIN in Figure 2. Given a current marketing instance $z_i = (\boldsymbol{x}_i, \boldsymbol{t}_i, y_i)$, the feature encoder module will encode non-treatment features $\boldsymbol{x}_i$ and treatment features $\boldsymbol{t}_i$ separately to obtain their respective embedding representations, i.e., $\mathbf{e}^x$ and $\mathbf{e}^t$. The embedded representation $\mathbf{e}^x$ will be fed into a self-interaction module with a self-attention network and multiple multilayer perceptrons, which computes the natural response of the instance when it is not treated, i.e., $\hat{y}_i(0)$. The embedded representations $\mathbf{e}^x$ and $\mathbf{e}^t$ will be fed into a treatment-aware interaction module to compute the ITE this instance has for a particular treatment, i.e., $\hat{\tau}_k(\boldsymbol{x}_i)$, where a treatment-aware attention network will model the interaction of $\mathbf{e}^x$ and $\mathbf{e}^t$. In addition, the estimated ITE will be combined with the previously predicted natural response to generate the response of this instance to a particular treatment, i.e., $\hat{y}_i(k) = \hat{y}_i(0) + \hat{\tau}_k(\boldsymbol{x}_i)$. The input of an intervention constraint module is the embedded representation $\mathbf{e}^{xt}$ containing interaction information obtained after going through the treatment-aware attention network, and the goal is to predict the group to which this instance belongs, i.e., $\hat{t}_{i0}$. The final optimization objective function of our EFIN can be expressed as follows,

$$\min_{\theta} \mathcal{L}_{EFIN} = \mathcal{L}_S + \mathcal{L}_T + \mathcal{L}_C + \lambda \|\theta\|, \tag{2}$$

where $\mathcal{L}_S$, $\mathcal{L}_T$, and $\mathcal{L}_C$ denote the losses for the self-interaction module, treatment-aware interaction module, and intervention constraint module, respectively, and λ and $\|\theta\|$ are the tradeoff parameter and the regularization terms.

4.2 Training

In this subsection, we describe each module in detail based on the training process.

4.2.1 The Feature Encoder Module. Given a current marketing instance $z_i = (\boldsymbol{x}_i, \boldsymbol{t}_i, y_i)$, unlike most existing works, we encode not only non-treatment features $\boldsymbol{x}_i$ but also treatment features $\boldsymbol{t}_i$ in this module. Taking the treatment feature as an example, for each continuous feature t_{ij} in the treatment features, we equip it with a shared fully-connected network for encoding. For the remaining features in the treatment features, i.e., the sparse features, we initialize an embedding table for each feature and obtain the embedding representation corresponding to the specific feature value through the *lookup* operation. This encoding process can be expressed as,

$$\mathbf{e}^t_{ij} = \begin{cases} \mathbf{W}_j * t_{ij} + \mathbf{b}_j, & t_{ij} \text{ is a continuous feature,} \\ lookup(\mathbf{E}_j, e_{ij}), & t_{ij} \text{ is a sparse feature,} \end{cases} \tag{3}$$

where $\mathbf{W}_j$ is a weight matrix, $\mathbf{b}_j$ is a bias vector, and $\mathbf{E}_j$ is an embedding table. Intuitively, continuous features in treatment features are more likely to reflect the correlation between different treatments, such as similar amounts and minimum consumption to be satisfied, so the different encoding of continuous features in Eq.(3) aims to preserve this property. For non-treatment features, we adopt a similar encoding process. Finally, we can obtain the corresponding embedding representation, i.e., $\mathbf{e}^x_i = \{\mathbf{e}^x_{i0}, \mathbf{e}^x_{i1}, \ldots, \mathbf{e}^x_{id_x}\}$ and $\mathbf{e}^t_i = \{\mathbf{e}^t_{i0}, \mathbf{e}^t_{i1}, \ldots, \mathbf{e}^t_{id_t}\}$.

4.2.2 The Self-interaction Module. In this module, we use the embedding representation $\mathbf{e}^x_i$ to model the natural response of each user in the control group, where information about the treatment is isolated to capture user-sensitive features in the natural situation. We use a self-attention network for self-interaction to better predict natural responses. Specifically, we have,

$$Q = K = V = \left(\mathbf{e}^x_{i0}; \mathbf{e}^x_{i1}; \ldots; \mathbf{e}^x_{i2}; \ldots; \mathbf{e}^x_{id_x}\right), \tag{4}$$

$$\overline{\mathbf{e}}^x_i = softmax(\frac{QK^T}{\sqrt{K_d}})V, \tag{5}$$

where K_d is the dimension of the output embedding, and $\overline{\mathbf{e}}^x_i = \{\overline{\mathbf{e}}^x_{i0}, \overline{\mathbf{e}}^x_{i1}, \ldots, \overline{\mathbf{e}}^x_{id_x}\}$. Next, we use a multilayer perceptron to predict natural responses,

$$\hat{y}_i(0) = \mathbf{W}_s * concat(\overline{\mathbf{e}}^x_i) + \mathbf{b}_s, \tag{6}$$

where $\mathbf{W}_s$ is a weight matrix and $\mathbf{b}_s$ is a bias vector. Based on the control group instances contained in the training set, the optimization objective of the self-interaction module is a supervised loss for natural responses,

$$\mathcal{L}_S = \mathcal{L}(\hat{y}_i(0), y_i(0)). \tag{7}$$

4.2.3 The Treatment-aware Interaction Module. In this module, we aim to use the embedding representations $\mathbf{e}^x_i$ and $\mathbf{e}^t_i$ to learn the responses of users in different treatment groups and to identify the corresponding sensitive features, where the treatment information will be used as inducements to achieve this goal. Note that this is different from the self-interaction module. Specifically, we first use a treatment-aware attention network to model the interaction between treatment features and non-treatment features, and use attention weights to describe the sensitivity of non-treatment features to a particular treatment,

$$\alpha^i_j = \text{Softmax}(\mathbf{W}^\top_{t0} \text{Relu}(\mathbf{W}_{t1}\mathbf{e}^t_i + \mathbf{W}_{t2}\mathbf{e}^x_{ij} + \mathbf{b}_{t2})), \tag{8}$$

$$\mathbf{e}^{xt}_i = \sum_{j=1}^{d_x} \alpha^i_j \mathbf{e}^x_{ij}, \tag{9}$$

where $\mathbf{W}_{t0}$, $\mathbf{W}_{t1}$ and $\mathbf{W}_{t2}$ are weight matrices and $\mathbf{b}_{t2}$ is a bias vector. Based on this embedded representation combined with interaction information, we then estimate the ITE of users in different treatment groups,

$$\hat{\tau}_k(\mathbf{x}_i) = \mathbf{W}_{t3} * \mathbf{e}^{xt}_i + \mathbf{b}_{s3}, \tag{10}$$

where $\mathbf{W}_{t3}$ is a weight matrix and $\mathbf{b}_{t3}$ is a bias vector. By combining the estimated ITE with the natural responses predicted by Eq.(6), we can obtain the responses of users in different treatment groups,

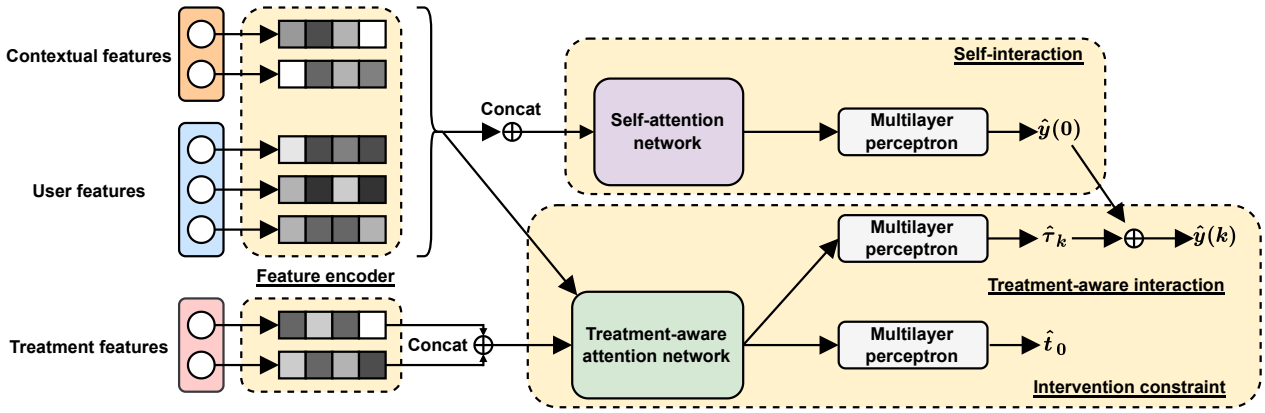

Figure 2: The architecture of the explicit feature interaction-aware uplift network (EFIN)

and construct the supervised loss of this module with instances of different treatment groups in the training set,

$$\hat{y}_i(k) = \hat{y}_i(0) + \hat{\tau}_k(\boldsymbol{x}_i). \tag{11}$$

$$\mathcal{L}_T = \mathcal{L}(\hat{y}_i(k), y_i(k)). \tag{12}$$

4.2.4 The Intervention Constraint Module. Since in an online marketing scenario, the assignment of different treatments is usually not random, and this means that the collected training set usually has a significant distribution difference between the control and treatment groups. As shown in Figure 3, since only one type of response in each group is available for supervised training, differences in distribution between groups will exacerbate significant differences in estimated ITE between groups, such as τ'_k and τ^*_k. Therefore, ignoring this difference may increase the difficulty of ITE estimation and impair accuracy. To alleviate this problem, we propose a simple but effective intervention constraint module. The idea behind this module is to increase the difficulty of guessing the corresponding group from the ITE distribution of different groups, that is, to achieve a trade-off between the two through mutual interference. Previous studies have shown that similarities in ITE distribution across groups are beneficial for uplift modeling [9]. Specifically, we use the embedding representation $\mathbf{e}_i^{xt}$, which is closely related to the uplift, to make predictions about which group this instance belongs to. We then train it with an inverse group label to generate perturbations as described above. This process can be expressed as,

$$\hat{t}_{i0} = \mathbf{W}_c * \mathbf{e}_i^{xt} + \mathbf{b}_c. \tag{13}$$

$$\mathcal{L}_C = \mathcal{L}(\hat{t}_{i0}, \bar{t}_{i0}). \tag{14}$$

Note that in the case of binary treatment, $\bar{t}_{i0}$ can directly take the opposite label. In the case of multi-valued treatment, a 0-1 mask vector needs to be generated from the original labels, and then the labels are negated.

4.2.5 The Uplift Prediction. After our EFIN training is complete, in the inference phase, we only need to use the treatment-aware interaction module to directly compute ITE, and then perform ranking and decision-making.

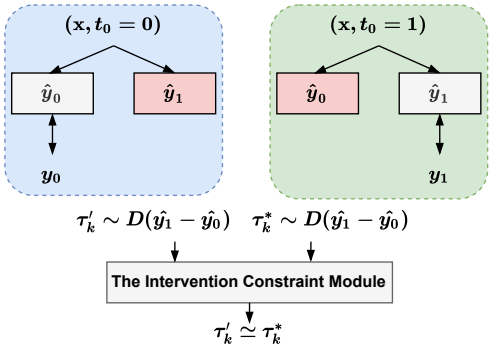

Figure 3: Illustration of the idea behind the intervention constraint module.

5 EMPIRICAL EVALUATIONS

In this section, we conduct experiments with the aim of answering the following three key questions.

- RQ1: How does our EFIN perform compared to the baselines?
- RQ2: What is the role of each module in our EFIN?
- RQ3: How effective is our EFIN in an online deployment?

5.1 Experimental Setup

5.1.1 Datasets. Following the settings of a previous work [16], we conduct experiments on two public datasets including CRITEO-UPLIFT [11] and EC-LIFT [16]. CRITEO-UPLIFT is a dataset open sourced by Criteo AI Labs for uplift modeling in a large-scale advertising scenario, which includes nearly 14 million instances, twelve continuous features, and binary treatments. EC-LIFT is a dataset of uplift modeling for different brands in a large-scale advertising scene, which is open sourced by Alimama. This dataset contains billions of instances, twenty-five discrete features and nine multi-valued features, and binary treatments. Due to the excessively large data scale, in order to facilitate training, we extracted about 40% of the instances from the original EC-LIFT dataset as the experimental dataset. The statistics of the two public datasets are shown in

Table 1. We randomly split the two dataset for training and testing with a ratio 8/2. Note that since the modeling of user features and contextual features is consistent and does not require a special distinction, we use all the features in the above datasets in our experiments. Furthermore, following the setting of previous work, we treat treatment as a binary feature. To comprehensively evaluate our EFIN, we also include a product dataset collected from two weeks of online coupon marketing scenarios for credit card repayments. This product dataset uses a total of more than 200 features, involves 2 million users and has 2 million instances, 90% of which are used for the training set and the rest for the test set. In particular, instead of binary treatments in the public datasets, seven treatment options are included in product dataset.

Table 1: Statistics of the two public datasets.

Dataset	CRITEO-UPLIFT	EC-LIFT
Size	13,979,592	196,084,380
Ratio of Treatment to Control	5.67:1	3.11:1
Average Visit Ratio	4.70%	3.25%
Relative Average Uplift	27.07%	464.46%
Average Uplift	1.03%	3.56%
Conversion Target	Visit	Visit

5.1.2 Evaluation Metrics. We evaluate uplift ranking performance by four widely used metrics, i.e., uplift score at first h percentile (LIFT@h), normalized area Under the uplift curve (AUUC), normalized area under the qini curve (QINI) and Weighted average uplift (WAU). We report the results with h set to 30. We use a standard python package *scikit-uplift*[1] to compute these metrics.

5.1.3 Baselines. To evaluate the effectiveness of our EFIN, we select a set of the most representative methods in neural network-based uplift modeling, including S-Learner [17], T-Learner [17], TarNet [30], CFRNet [30], DragonNet [31], GANITE [35], CEVAE [21], SNet [8], FlexTENet [9], EUEN [16] and DESCN [39].

5.1.4 Implementation Details. We implement all baselines and our EFIN in PyTorch 1.13[2]. We use an AdamW optimizer[3] and set the maximum number of iterations to 20. To search for the best hyperparameters, we use QINI as a primary evaluation metric. We also adopt an early stopping mechanism with a patience of 5 to avoid overfitting to the training set. Furthermore, we use the hyperparameter search library *Optuna*[4] to accelerate the tuning process. The range of the values of the hyper-parameters are shown in Table 2.

5.2 RQ1: Performance Comparison

We report the comparison results on two public datasets in Table 3. From the results in Table 3, we can have the following observations: 1) T-learner significantly outperform S-learner, even some baselines using more complex network architectures. This may mean that in online marketing scenarios with numerous features, uplift modeling is more difficult than traditional ITE estimation, especially the sensitive features of users need to be more accurately identified. In particular, we can observe that on EC-LIFT with a large number of high-dimensional sparse features, most baselines no longer have an advantage over the T-learner. 2) By designing some more flexible or complex architectures as estimators of user responses, FlexTENet, SNet, EUEN and DESCN perform better than other baselines. But again, their advantage over EC-LIFT shrinks. This means that other architectural changes may not yield much gain without taking feature interactions into account. 3) Unlike other baselines, our EFIN consistently outperforms all baselines in most cases except slightly weaker than DESCN on WAU. Since we use QINI as the main metric in the hyperparameter search, there may be some fluctuations in other metrics, and we can find that our EFIN has a large improvement on QINI. Furthermore, our EFIN is also able to maintain the performance advantage on EC-LIFT, benefiting from the explicit modeling of treatment features and feature interactions.

Table 2: Hyper-parameters and their values tuned in the experiments.

Name	Range	Functionality
$rank$	$\{2^5, 2^6, 2^7\}$	Embedded dimension
bs	$\{2^8, 2^9, 2^{10}, 2^{11}\}$	Batch size
lr	$\{1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}\}$	Learning rate
λ	$\{1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}\}$	Loss weighting

Next, we report the comparison results on the product dataset in Table 4. Since most of the baselines are usually only applied to binary treatment scenarios, to evaluate them on the product dataset with multi-valued treatments, we first extend them reasonably, such as the network architecture changing from two-head to multi-head. Note that since the distribution estimation in CEVAE is difficult to directly extend to the multi-head architecture, we do not report its results on the product dataset. After the expansion is complete, we retrain all the methods using the same search range as in Table 2. Note that when evaluating, we need to treat multi-valued treatments as multiple binary treatments to obtain individual metrics for each treatment, and finally report the averaged results. From the results in Table 4, we have the following observations: 1) meta-learner-based methods (S-Learner and T-Learner) are still relatively stable and have suboptimal results in multi-valued treatments scenarios. 2) the baselines considering a shared architecture suffer from a performance bottleneck, where the shared part may cause learning shocks due to too many and significantly different treatment groups. 3) similarly, since our EFIN exploits treatment features and feature interactions explicitly, on the product dataset it still retains the ability to mine for each user its sensitive features associated with different treatments. Combined with the results on two public datasets and one product dataset, this both validates the effectiveness of our EFIN, especially considering treatment features and feature interactions explicitly in the uplift modeling.

[1] https://www.uplift-modeling.com/en/latest/
[2] https://pytorch.org/
[3] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html
[4] https://optuna.org/

Table 3: Results on two public datasets, where the best and second best results are marked in bold and underlined, respectively. Note that * indicates a significance level of $p \leq 0.05$ based on two-sample t-test between our method and the best baseline.

Dataset	CRITEO-UPLIFT				EC-LIFT			
Metrics	LIFT@30	QINI	AUUC	WAU	LIFT@30	QINI	AUUC	WAU
S-Learner	0.0328	0.0857	0.0332	0.0092	0.0080	0.0414	0.0073	0.0031
T-Learner	0.0425	0.1083	0.0430	0.0093	0.0086	0.0440	0.0079	0.0032
TarNet	0.0339	0.1027	0.0406	0.0087	0.0081	0.0422	0.0076	0.0031
CFRNet	0.0379	0.1052	0.0414	0.0101	0.0087	0.0422	0.0078	0.0031
DragonNet	<u>0.0464</u>	0.1096	0.0437	0.0093	<u>0.0096</u>	<u>0.0459</u>	<u>0.0092</u>	0.0033
GANITE	0.0447	<u>0.1170</u>	<u>0.0468</u>	0.0101	0.0080	0.0409	0.0068	0.0029
CEVAE	0.0365	0.0951	0.0375	0.0106	0.0077	0.0373	0.0068	0.0031
FlexTENet	0.0448	0.1108	0.0441	0.0093	0.0084	0.0435	0.0078	0.0031
SNet	0.0442	0.1112	0.0442	0.0083	0.0084	0.0441	0.0079	0.0032
EUEN	0.0425	0.1153	0.0457	0.0108	0.0090	0.0446	0.0084	0.0033
DESCN	0.0456	0.1129	0.0455	**0.0131***	0.0082	0.0435	0.0075	<u>0.0034</u>
EFIN	**0.0468***	**0.1285***	**0.0514***	<u>0.0122</u>	**0.0100***	**0.0468***	**0.0097***	**0.0034**

Table 4: Results on a product dataset, where the best and second best results are marked in bold and underlined, respectively. Note that * indicates a significance level of $p \leq 0.05$ based on two-sample t-test between our method and the best baseline.

Dataset	Product	
Metrics	Average QINI	Average AUUC
S-Learner	0.0155	**0.0094***
T-Learner	<u>0.0158</u>	0.0034
TarNet	0.0118	0.0001
CFRNet	0.0110	0.0066
DragonNet	0.0136	0.0004
GANITE	0.0101	0.0047
CEVAE	-	-
FlexTENet	0.0143	0.0076
SNet	0.0122	0.0064
EUEN	0.0088	0.0003
DESCN	0.0128	0.0003
EFIN	**0.0172***	<u>0.0085</u>

5.3 RQ2: Ablation Study of EFIN

Moreover, we conduct ablation studies of our EFIN to analyze the role played by each proposed module. We sequentially removed the three core modules individually, i.e., the self-interaction module, the treatment-aware interaction module, and the intervention constraint module. The results are shown in Table 5. From the results in Table 5, we can find that removing any module will bring performance degradation. This verifies the validity of each module design in our EFIN. That is, the intervention constraint module can make distribution adjustments to the data collected by non-random treatment assignment, and the self-interaction module and treatment-aware interaction module can capture different sensitive features of the users in natural and treatment situations, respectively.

5.4 RQ3: Results of the Online Deployment

To further evaluate the performance, we deploy our EFIN on credit card payment scenario in FiT Tencent, which is one of the large-scale online financial platform in China.

5.4.1 System Overview & Scenario Description. The scenario is illustrated in Figure 4. Marketing in this scenario needs to launch different campaigns for different customer groups to motivate more users to pay the credit card bill within this platform. The treatments are promoted to some group of users once user tends to pay credit card bill on the platform. There are various types of coupons according to the constraints on bill amount in this scenario. Specifically, there are some small denomination coupons without a minimum amount requirement, and some higher denomination coupons requires minimum amount, thus the final number of treatments in this scenario is set as 7. The high-level architecture can be show in Figure 5. The user behaviors are pulled from data sources (some storage cluster) to generate features using Apache Spark[5] and the candidate of treatments is coupons with various amount. The uplift model will score the uplift value for every user on each coupon. Finally, the promoter platform will further deliver coupons to user group with some resource constraints. Notice that our work focus on how to improve the performance of uplift model, which is the key component in the whole system.

5.4.2 Online Experimental Results. To conduct online A/B experiment, we divide two sets of online traffic that do not affect each other, which involved hundreds of millions of users. The existing

[5]https://spark.apache.org/

Table 5: Ablation study of our EFIN on CRITEO-UPLIFT.

Dataset	CRITEO-UPLIFT			
Metrics	LIFT@30	QINI	AUUC	WAU
w/o Self-Interactive Module	0.0467	0.1266	0.0506	0.0104
w/o Treatment-aware Interaction	**0.0484**	0.1254	0.0501	0.0107
w/o Intervention Constraint Module	0.0442	0.1172	0.0465	0.0109
EFIN	0.0468	**0.1285**	**0.0514**	**0.0122**

Figure 4: The illustration of credit card scenario.

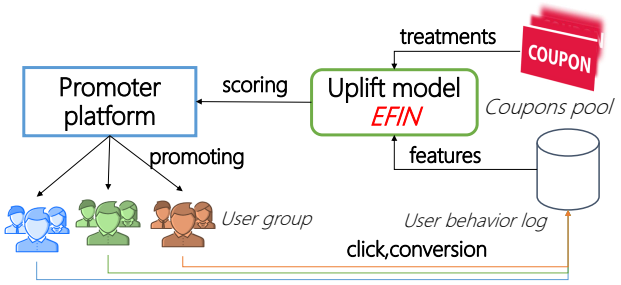

Figure 5: Overview of the promotion system in FiT Tencent.

baseline on the online platform is a multi-head extended T-learner, where each estimator employs XGBoost [6] for computation. The baseline model and our EFIN are served for each independent online traffic for one month. As to evaluate the performance, we use two important online metrics: the marketing return on investment (**ROI**) and the number of monthly active users (**MAU**). Table 6 reports the relative improvements over the baseline. From the results in Table 6, we can find that our EFIN improves ROI and MAU by 10% and 8% compared to the baseline, respectively. This means that our EFIN can maintain a stable performance advantage over a period of time, and can indeed accurately capture the sensitive characteristics of different users to perform a reasonable treatment assignment.

Table 6: Results of our EFIN in an online deployment.

Metrics	ROI	MAU
Base (T-Learner + XGBoost)	0.0%	0.0%
EFIN	+10%	+8%

6 CONCLUSIONS AND FUTURE WORK

In this paper, in order to address the underutilization of treatment features and feature interactions that exists in most existing uplift modeling methods, we propose an explicit feature interaction-aware uplift network (EFIN). Our EFIN consists of four modules, where a feature encoder module is used to encode all features, a self-interaction model aims to accurately model a user's natural response using non-treatment features while isolating treatment information, a treatment-aware interaction module utilizes both treatment features and non-treatment features, and accurately models a user's uplift and response to different treatments through their interactions, and an intervention constraint module is designed to adjust for the distributional differences in the control and treatment

groups to make our EFIN more robust across different scenarios. Finally, we conduct extensive offline and online evaluations and the results validate the effectiveness of our EFIN.

For future work, we plan to explore and analyze the effectiveness of more feature interaction architectures in uplift modeling. It is also a promising question how to make uplift modeling benefit more from the treatment feature and its interaction with non-treatment features. In addition, we are also interested in considering and solving some of the more complex uplift modeling scenarios, such as considering necessary constraints like net profit, and modeling a user's response changes to different treatments based on a dynamic perspective.

ACKNOWLEDGMENTS

We thank the support of National Natural Science Foundation of China Nos. 61836005, 62272315 and 62172283.

REFERENCES

[1] Meng Ai, Biao Li, Heyang Gong, Qingwei Yu, Shengjie Xue, Yuan Zhang, Yunzhou Zhang, and Peng Jiang. 2022. LBCF: A large-scale budget-constrained causal forest algorithm. In *Proceedings of the ACM Web Conference 2022*. 2310–2319.

[2] Javier Albert and Dmitri Goldenberg. 2021. E-commerce promotions personalization via online multiple-choice knapsack with uplift modeling. *arXiv preprint arXiv:2108.13298* (2021).

[3] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.

[4] Heejung Bang and James M Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61, 4 (2005), 962–973.

[5] Artem Betlei, Eustache Diemert, and Massih-Reza Amini. 2021. Uplift modeling with generalization guarantees. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 55–65.

[6] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[7] Xuanying Chen, Zhining Liu, Li Yu, Liuyi Yao, Wenpeng Zhang, Yi Dong, Lihong Gu, Xiaodong Zeng, Yize Tan, and Jinjie Gu. 2022. Imbalance-aware uplift modeling for observational data. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*. 6313–6321.

[8] Alicia Curth and Mihaela van der Schaar. 2021. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*. 1810–1818.

[9] Alicia Curth and Mihaela van der Schaar. 2021. On inductive biases for heterogeneous treatment effect estimation. *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 15883–15894.

[10] Floris Devriendt, Jente Van Belle, Tias Guns, and Wouter Verbeke. 2020. Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering* 34, 10 (2020), 4888–4904.

[11] Eustache Diemert, Artem Betlei, Christophe Renaudin, Massih-Reza Amini, Théophane Gregoir, and Thibaud Rahier. 2021. A large scale benchmark for individual treatment effect prediction and uplift modeling. *arXiv preprint arXiv:2111.10106* (2021).

[12] Dmitri Goldenberg, Javier Albert, Lucas Bernardi, and Pablo Estevez. 2020. Free lunch! retrospective uplift modeling for dynamic promotions recommendation within roi constraints. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 486–491.

[13] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 1725–1731.

[14] Xiaofeng He, Guoqiang Xu, Cunxiang Yin, Zhongyu Wei, Yuncong Li, Yancheng He, and Jing Cai. 2022. Causal enhanced uplift model. In *Proceedings of the 26th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 119–131.

[15] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. Field-aware factorization machines for CTR prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 43–50.

[16] Wenwei Ke, Chuanren Liu, Xiangfu Shi, Yiqiao Dai, S Yu Philip, and Xiaoqiang Zhu. 2021. Addressing exposure bias in uplift modeling for large-scale online advertising. In *Proceedings of the 2021 IEEE International Conference on Data Mining*. 1156–1161.

[17] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 4156–4165.

[18] Zekun Li, Zeyu Cui, Shu Wu, Xiaoyu Zhang, and Liang Wang. 2019. Fi-GNN: Modeling feature interactions via graph neural networks for ctr prediction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 539–548.

[19] Ying-Chun Lin, Chi-Hsuan Huang, Chu-Cheng Hsieh, Yu-Chen Shu, and Kun-Ta Chuang. 2017. Monetary discount strategies for real-time promotion campaign. In *Proceedings of the ACM Web Conference 2017*. 1123–1132.

[20] Dugang Liu, Mingkai He, Jinwei Luo, Jiangxu Lin, Meng Wang, Xiaolian Zhang, Weike Pan, and Zhong Ming. 2022. User-event graph embedding learning for context-aware recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1051–1059.

[21] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6449–6459.

[22] Fuyuan Lyu, Xing Tang, Huifeng Guo, Ruiming Tang, Xiuqiang He, Rui Zhang, and Xue Liu. 2022. Memorize, factorize, or be naive: Learning optimal feature interaction methods for CTR prediction. In *2022 IEEE 38th International Conference on Data Engineering*.

[23] Fuyuan Lyu, Xing Tang, Dugang Liu, Liang Chen, Xiuqiang He, and Xue Liu. 2023. Optimizing feature set for click-through rate prediction. In *Proceedings of the ACM Web Conference 2023*. 3386–3395.

[24] X Nie and S Wager. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108, 2 (2021), 299–319.

[25] Nicholas J Radcliffe and Patrick D Surry. 2011. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions* (2011), 1–33.

[26] Steffen Rendle, Zeno Gantner, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2011. Fast context-aware recommendations with factorization machines. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 635–644.

[27] Thomas Reutterer, Andreas Mild, Martin Natter, and Alfred Taudes. 2006. A dynamic segmentation approach for targeting and customizing direct marketing campaigns. *Journal of Interactive Marketing* 20, 3-4 (2006), 43–57.

[28] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[29] Yuta Saito, Hayato Sakata, and Kazuhide Nakata. 2020. Cost-effective and stable policy optimization algorithm for uplift modeling with multiple treatments. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. 406–414.

[30] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. 3076–3085.

[31] Claudia Shi, David M Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 2507–2517.

[32] Shu Wan, Chen Zheng, Zhonggen Sun, Mengfan Xu, Xiaoqing Yang, Hongtu Zhu, and Jiecheng Guo. 2022. GCF: Generalized causal forest for heterogeneous treatment effect estimation in online marketplace. *arXiv preprint arXiv:2203.10975* (2022).

[33] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. DCN V2: Improved deep and cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the ACM Web Conference 2021*. 1785–1797.

[34] Guoqiang Xu, Cunxiang Yin, Yuchen Zhang, Yuncong Li, Yancheng He, Jing Cai, and Zhongyu Wei. 2022. Learning discriminative representation base on attention for uplift. In *Proceedings of the 26th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 200–211.

[35] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *Proceedings of the 6th International Conference on Learning Representations*.

[36] Weijia Zhang, Jiuyong Li, and Lin Liu. 2021. A unified survey of treatment effect heterogeneity modelling and uplift modelling. *Comput. Surveys* 54, 8 (2021), 1–36.

[37] Kui Zhao, Junhao Hua, Ling Yan, Qi Zhang, Huan Xu, and Cheng Yang. 2019. A unified framework for marketing budget allocation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1820–1830.

[38] Yan Zhao, Xiao Fang, and David Simchi-Levi. 2017. Uplift modeling with multiple treatments and general response types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. 588–596.

[39] Kailiang Zhong, Fengtong Xiao, Yan Ren, Yaorong Liang, Wenqing Yao, Xiaofeng Yang, and Ling Cen. 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.