Rankability-enhanced Revenue Uplift Modeling Framework for Online Marketing

Bowei He*
City University of Hong Kong
Hong Kong SAR
boweihe2-c@my.cityu.edu.hk

Yunpeng Weng
FiT, Tencent
Shenzhen, China
wengyp@mail3.sysu.edu.cn

Xing Tang
FiT, Tencent
Shenzhen, China
xing.tang@hotmail.com

Ziqiang Cui
City University of Hong Kong
Hong Kong SAR
ziqiang.cui@my.cityu.edu.hk

Zexu Sun
Renmin University of China
Beijing, China
sunzexu21@ruc.edu.cn

Liang Chen
FiT, Tencent
Shenzhen, China
leocchen@tencent.com

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

Chen Ma†
City University of Hong Kong
Hong Kong SAR
chenma@cityu.edu.hk

Abstract

Uplift modeling has been widely employed in online marketing by predicting the response difference between the treatment and control groups, so as to identify the sensitive individuals toward interventions like coupons or discounts. Compared with traditional *conversion uplift modeling*, *revenue uplift modeling* exhibits higher potential due to its direct connection with the corporate income. However, previous works can hardly handle the continuous long-tail response distribution in revenue uplift modeling. Moreover, they have neglected to optimize the uplift ranking among different individuals, which is actually the core of uplift modeling. To address such issues, in this paper, we first utilize the zero-inflated lognormal (ZILN) loss to regress the responses and customize the corresponding modeling network, which can be adapted to different existing uplift models. Then, we study the ranking-related uplift modeling error from the theoretical perspective and propose two tighter error bounds as the additional loss terms to the conventional response regression loss. Finally, we directly model the uplift ranking error for the entire population with a listwise uplift ranking loss. The experiment results on offline public and industrial datasets validate the effectiveness of our method for revenue uplift modeling. Furthermore, we conduct large-scale experiments on a prominent online fintech marketing platform, Tencent FiT, which further demonstrates the superiority of our method in real-world applications.

*Work done as an intern in FiT, Tencent
†Corresponding author

CCS Concepts

Keywords

1 Introduction

Uplift modeling [17], aiming to predict the expected difference between the treatment and control response, has been widely adopted to identify the individuals among a targeted population who can react positively to a particular intervention. Recently, this technique has been successfully deployed in many scenarios, like healthcare, finance transactions, and online marketing [21, 29, 37]. Generally, the uplift modeling in online marketing can be divided into two categories [15]: *conversion uplift modeling* and *revenue uplift modeling*. The latter one is the focus of this work considering its higher application value and broader application scenarios.

Several methods have been proposed to perform uplift modeling to infer the causal effect of a specific treatment, such as price discount, repayment incentive, or coupon delivery. The most common one is the Randomized Controlled Trial (RCT) [12, 31] like the A/B test for marketing, where each individual is randomly assigned to the treatment group or the control group, thereby being independent of other covariate features. However, RCT experiments are often expensive, time-consuming, and even harmful to platforms in many applications. Even worse, RCT can only estimate the average uplift effect in the whole population, far from the individual uplift effect estimation, which is the pursuit of the current online marketing. Therefore, most recent works focus on learning an uplift

model from the experimental data and then directly deploying it to estimate the uplift effect for studied individuals. Among them, the meta-learner methods [22], like S-Learner and T-Learner are the pioneer works seeking to estimate the uplift effect of personalized treatments on different individuals. However, they can be easily manipulated by the sample imbalance issue between the treatment and control group. Several tree-based methods [11, 28], like Causal Forest, are proposed to address this issue by adapting the splitting criteria of conventional decision trees. Representation learning methods [29, 30] are another type of dominant approach to this issue with the development of deep neural networks especially in recent years. They mainly balance the representation distributions of individuals in the treatment group and control group.
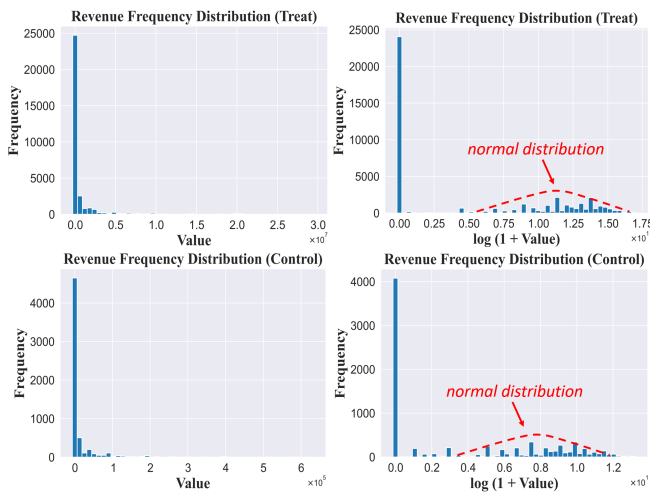

Figure 1: The histogram for the revenue frequency distribution of the A/B test in our deployment platform's online fintech application scenario. Left: raw revenue values; Right: revenue values after the logarithmic transformation.

Though achieving acceptable performance in some idealized experiment environments or synthetic datasets, there are still some avenues to further improve the above methods. First, few of them focus on revenue uplift modeling, where the response is a *continuous variable* rather than a simple binary one. This obviously increases the difficulty of precise uplift modeling. Meanwhile, this type of uplift modeling problem often has higher practical value though less explored, considering that practitioners hope to increase the return (continuous response) on marketing investments in many cases. Second, most of the existing methods [10, 28–30] only work well on synthetic datasets, where the range of responses is limited, which is far from the real production environments, especially in finance scenarios. In such cases, the expected response, like the mutual fund sales revenue, can range from zero/several dollars to several million or even billion dollars. If directly adopting previous uplift methods to such tasks, some extreme data points can bring huge obstacles to model learning, thus the performance can hardly be satisfactory. Besides, the distribution of revenue response can be extremely imbalanced. In the above case, the sales revenue of most users hardly exceeds 10,000 USD, while only a very tiny proportion of users may spend over 500,000 USD. These two points can be concluded as

the *long-tail distribution* challenge, which is pretty common in online applications, like the one shown in Figure 1. Last but not least, almost all previous methods only care about the accuracy of the uplift effect prediction for different individuals, neglecting the ranking accuracy among them. That is, the individuals with higher true uplift values should obtain higher predicted uplift values compared with those lower-uplift value individuals, though the uplift value prediction is not so precise. In fact, the *rankability of uplift models* is a more serious and meaningful problem in many real-world applications, because the ultimate purpose of utilizing the uplift models is to identify the individuals who are more susceptible to the treatment and then intervene with them differently according to the ranking of their uplift values. Moreover, this problem is more challenging in scenarios with *long-tail* responses. Because in the learning phase, the model will mainly pay attention to individuals with high responses, and ignore the prediction accuracy of the individuals with lower responses. Thus, the uplift ranking confusion for these latter individuals can be exacerbated.

To address the aforementioned challenges, we propose a **R**anking-**E**nhanced **R**evenue **U**plift **M**odeling (**RERUM**) framework. First, to overcome the continuous value and long-tail distribution challenges, we replace the conventional Mean-Squared-Error (MSE) loss with a ZILN loss for revenue response regression in treatment and control groups. Meanwhile, an accompanying regression network framework is proposed to adapt to ZILN loss. Second, we make a theoretical analysis of the uplift ranking error and propose more stringent error bounds as additional revenue response losses, which can help enhance the model's uplift ranking ability from the revenue response ranking perspective. Finally, we directly consider the uplift ranking among different individuals in the targeted population and propose a listwise ranking loss to explicitly optimize the uplift model's rankability.

To summarize, the main contributions of this work are as follows:

- To tackle the long-tail issue in revenue response regression, we propose the ZILN loss and its corresponding model framework.
- To mitigate the uplift modeling error, we perform a theoretical analysis and employ two tighter error bounds as response ranking losses to augment the above regression loss.
- To directly enhance the rankability of the uplift model over the whole population, we provide a listwise uplift ranking loss.
- To demonstrate the effectiveness of our method, we conduct extensive offline experiments on both public and industrial datasets. Additionally, large-scale online experiments on a fintech marketing platform with over 400 million users are also performed.

2 Related Works

Uplift Modeling. The uplift modeling in online marketing is typically comprised of two types of problems: *conversion uplift modeling* and *revenue uplift modeling*. In conversion uplift modeling [3, 4, 11, 21, 22], the response label is a simple 0/1 binary variable indicating if the individual purchases the goods or takes a specific action, like clicking the advertisement. Most of the previous research concentrates on this topic. Though achieving accepting performance in some scenarios, many of them [25, 29, 30, 34] are dedicated to the uplift value prediction, neglecting the importance of uplift ranking among the population, which is originally the core point. Revenue uplift modeling [15, 16, 37], however, differs from

the traditional conversion uplift modeling on that its response label is a continuous variable with unlimited value range and irregular distribution, like the paying expense. This obviously introduces extra challenges for accurate uplift modeling. Unfortunately, few of works focus on the revenue uplift modeling problem, especially the rankability of revenue uplift models, though its higher application value and wider application scenarios in various online marketing platforms. In this paper, we dive into the rankability-enhanced revenue uplift modeling framework and explore its application in real-world scenarios.

Learning to Rank. Learning to rank (LTR) has raised growing attention in recent decades and have been deployed in many different applications. The *pointwise* methods like Mcrank [24] are first proposed to solve this problem by transforming it to a regression or classification task on each single object. The models are trained to predict how relevant an object is for a given query. The core issue of this approach is that there exists the deviation between its optimization objective and the original ranking task. *Pairwise* methods are an alternative solution which takes object pairs into consideration and transforms the ranking into a binary classification task. In other words, the models need to judge whether the order of each pair is wrong relative to the ground truth. The representative works in this line include: RankSVM [20], Rankboost [14], RankNet [7], LambdaRank [6], and LambdaMART [8, 19]. *Pairwise* methods have more promising results in practice because they come closer to the nature of ranking than *pointwise* ones. However, they also possess a significant drawback, i.e., ignoring the discrepancy among different pairs. To address this issue, *listwise* methods like Adarank [36], ListNet [9], ListMLE [23, 35] have been proposed as a direct solution and consider all objects simultaneously. Nevertheless, most previous related literature is limited to the information retrieval tasks, like the document retrieval. Its application in causal effect estimation, especially the uplift modeling in online marketing is barely unexplored, though its great potential for improving the uplift model's ranking ability, which is the main focus of this work.

3 Preliminaries

We follow the Neyman-Rubin potential outcome framework [27] to formulate the revenue uplift modeling problem. In detail, we have the observed sample set $\mathcal{D} = \{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^{n}$. For each individual $i \in \mathcal{I}$ ($|\mathcal{I}| = n$), $y_i \in \mathcal{Y} \subset \mathbb{R}$ is a continuous outcome (response), $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of covariates, and $t_i \in \mathcal{T} = \{0, 1\}$ denotes the treatment (with $t_i = 0$ as the control) intervened on individual i. Y_i, X_i, and T_i are corresponding random variables, respectively. Furthermore, Y, X, and T are the general random variables that indicate the response, covariate, and treatment, respectively, regardless of the specific individual. Thus, the uplift effect τ_i of the treatment on individual i is defined as:

$$\tau_i = y_i^1 - y_i^0, \tag{1}$$

where y_i^1 and y_i^0 represent the potential responses under the corresponding treatment and control, respectively. Note that the whole population can be divided to two groups according to their received treatment: treatment group $\mathcal{D}^t = \{(\mathbf{x}_i, t_i, y_i)|(\mathbf{x}_i, t_i, y_i) \in \mathcal{D}, t_i = 1\}$, and control group $\mathcal{D}^c = \{(\mathbf{x}_i, t_i, y_i)|(\mathbf{x}_i, t_i, y_i) \in \mathcal{D}, t_i = 0\}$. $\mathbf{x}^{t/c}$ and $X^{t/c}$ represent the covariate vector instances of users in the

treatment/control group and their corresponding random variable, respectively. However, due to the non-simultaneous observability of potential responses, the uplift effect in the Neyman-Rubin potential outcome framework can hardly be accessed directly. An empirical alternative is the Conditional Average Treatment Effect (CATE) [2], which measures the response difference between the corresponding treatment and the control, conditioned on the observed covariates. In this paper, following previous works [4, 26], we estimate the CATE from the statistical perspective as the uplift effect, that is:

$$\tau(\mathbf{x}_i) = \mathbb{E}[Y_i|X_i = \mathbf{x}_i, T_i = 1] - \mathbb{E}[Y_i|X_i = \mathbf{x}_i, T_i = 0]. \tag{2}$$

It should be noted that when the uplift is estimated perfectly, the rankability of the model is also maximized.

4 Methodology

In this section, we first introduce the uplift estimation model framework based on the previous related models and the newly designed zero-inflated lognormal loss for the revenue response regression. Second, we theoretically analyze the uplift modeling error from the ranking perspective and derive the tighter upper error bounds as the response ranking learning objectives. Then, motivated by the listwise ranking research, we directly model the uplift ranking among different individuals in the targeted population. Finally, we conclude the overall training objective for rankability-enhanced revenue uplift modeling.

4.1 Uplift Estimation Model Framework

According to the uplift effect formulation (Eq. 2) in Sec. 3, conducting the accurate revenue response regression is undoubtedly crucial to the uplift modeling. However, previous uplift models [25, 29, 30, 34] mainly rely on the MSE loss to learn response models which can hardly adapt to the continuous long-tail distribution in revenue uplift scenario, considering its sensitivity to outliers. To better achieve the goal of accurate response regression, we first study the response distribution. From Figure 1, we have two observations: 1) small values, especially zeros, occupy the main frequency of the revenue response distribution, and 2) after the logarithmic transformation, except for such small (zero) values, the rest of the data roughly follows a normal distribution.

Motivated by the above observations, we refer to the related research in customer lifetime value research [33] and propose the zero-inflated lognormal (ZILN) loss for response regression instead of the conventional MSE loss. In detail, based on the previous observation 1), the zero-inflation phenomenon, we first design a cross-entropy loss with the probability p as the corresponding *purchasing propensity* loss to help determine if the individual will spend money. Only the individuals that do spend money (payers) will engage into the computation of the following *payer expense* loss. In this way, we can avoid the interference of such zero-value individuals in the positive revenue response regression. Furthermore, based on the above observation 2), we utilize the lognormal distribution with mean μ and standard deviation σ to model the positive revenue response from such payers and correspondingly derive its negative log-likelihood as the *payer expense* loss for non-zero response regression. In this lognormal loss, we conduct a logarithmic transformation to reduce the skewness of the response distribution. The
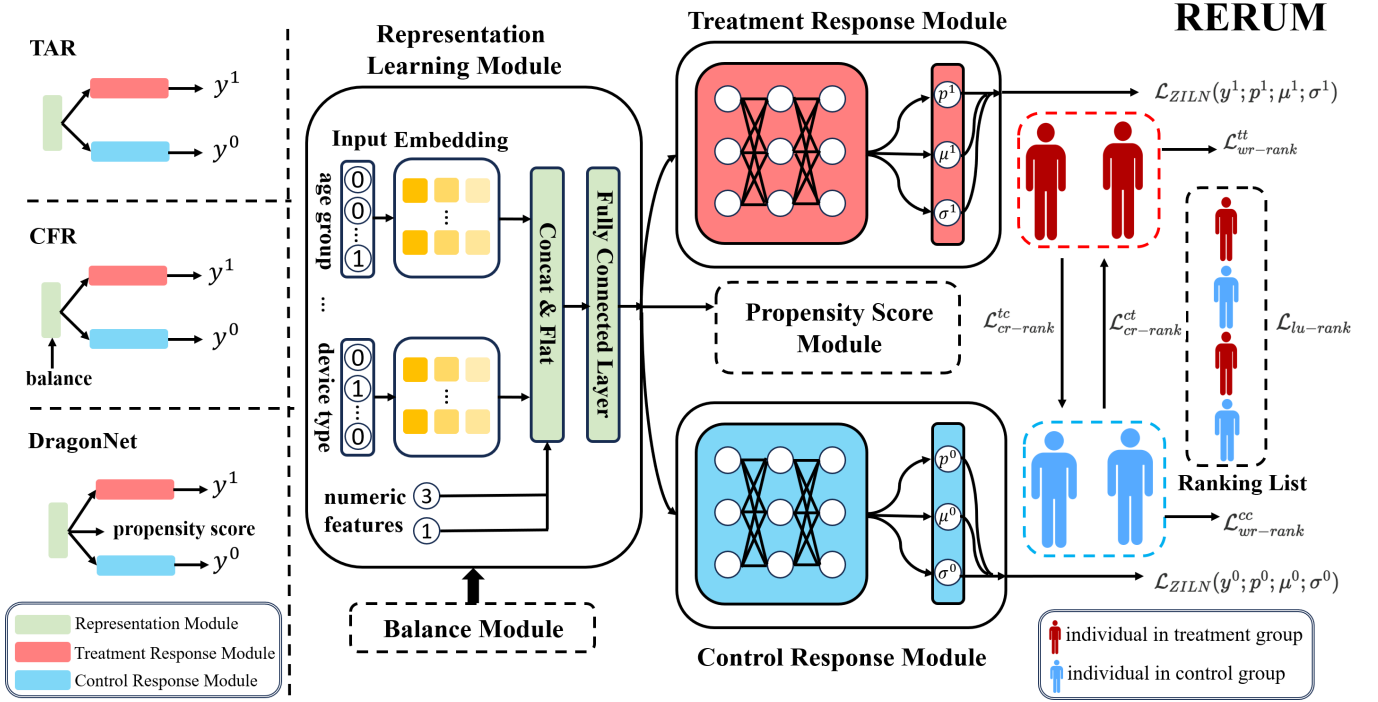
Figure 2: The overall framework of uplift estimation model. The modules in the dotted rectangular boxes are optional according to the specific base model selection. The original versions of base models take MSE as the response regression loss.

mathematical form of our ZILN loss is as follows:

$$\mathcal{L}_{ZILN}(y; p; \mu; \sigma) = \mathcal{L}_{purchasing\ propensity} + \mathcal{L}_{payer\ expense}$$
$$= \mathcal{L}_{CrossEntropy}(\mathbb{1}(y > 0); p) + \mathbb{1}(y > 0)\mathcal{L}_{Lognormal}(y; \mu; \sigma). \tag{3}$$

Furthermore, in this equation, we have

$$\mathcal{L}_{CrossEntropy}(\mathbb{1}(y > 0); p) = -\mathbb{1}(y = 0)\log(1 - p) - \mathbb{1}(y > 0)\log p,$$

$$\mathcal{L}_{Lognormal}(y; \mu; \sigma) = \log(y\sigma\sqrt{2\pi}) + \frac{(\log y - \mu)^2}{2\sigma^2}. \tag{4}$$

It should be mentioned that the ZILN loss can not only facilitate the agreement between the predicted response and the true response, but also help discriminate the ranking among the responses [33], which is beneficial to the response ranking learning module in the later Sec. 4.3.

To match the realization of the above ZILN loss, we customize the uplift estimation model framework based on the previous related uplift models shown on the left side of Figure 2. As shown in the right side of Figure 2, our newly designed framework includes the representation learning module, treatment response module, control response module, etc. For a given individual, the treatment/control response module will take the corresponding representation vector from the representation learning module to predict p^1, μ^1, σ^1 or p^0, μ^0, σ^0, respectively, which will be further utilized to obtain regression loss via Eq. 3. Note our proposed framework is agnostic to the specific uplift model choice; any previously mentioned base models in Sec. 2 can be integrated with our framework. In our following implementations, we mainly adopt the deep learning-based models as the workhorse due to their strong ability to capture the

complex and nonlinear relations between covariates and responses, as well as the competitive performance proved by previous works.

4.2 Uplift Modeling Error Analysis

To restore the correct ranking among the τ_i for different individuals, an intuitive approach is to ensure that the predicted uplift distance $\hat{\tau}_i - \hat{\tau}_j$ between each possible individual pair (i, j) is as close as possible to their true uplift distance $\tau_i - \tau_j$. Therefore, in the following part, we focus on the uplift distance prediction error $|(\hat{\tau}_i - \hat{\tau}_j) - (\tau_i - \tau_j)|$ and theoretically prove that conventional MSE loss for response regression is a loose error bound which can hardly capture the true uplift distance between individual pairs.

PROOF. Take randomly selected individuals i and j as an example. Firstly, according to the definition of the uplift effect, we have the following equations:

$$\hat{\tau}_i - \hat{\tau}_j = (\hat{y}_i^1 - \hat{y}_i^0) - (\hat{y}_j^1 - \hat{y}_j^0) = (\hat{y}_i^1 - \hat{y}_j^1) - (\hat{y}_i^0 - \hat{y}_j^0),$$
$$\tau_i - \tau_j = (y_i^1 - y_i^0) - (y_j^1 - y_j^0) = (y_i^1 - y_j^1) - (y_i^0 - y_j^0). \tag{5}$$

Thus, we can obtain the discrepancy between the predicted pairwise uplift distance and the true pairwise uplift distance for individual i and j as follows:

$$|(\hat{\tau}_i - \hat{\tau}_j) - (\tau_i - \tau_j)|$$
$$= |((\hat{y}_i^1 - \hat{y}_j^1) - (\hat{y}_i^0 - \hat{y}_j^0)) - ((y_i^1 - y_j^1) - (y_i^0 - y_j^0))|$$
$$= |((\hat{y}_i^1 - \hat{y}_j^1) - (y_i^1 - y_j^1)) - ((\hat{y}_i^0 - \hat{y}_j^0) - (y_i^0 - y_j^0))|$$
$$\leq |(\hat{y}_i^1 - \hat{y}_j^1) - (y_i^1 - y_j^1)| + |(\hat{y}_i^0 - \hat{y}_j^0) - (y_i^0 - y_j^0)|. \tag{6}$$

Based on the above derivation, we have:

$$
\begin{aligned}
&|(\hat{\tau}_i - \hat{\tau}_j) - (\tau_i - \tau_j)| \\
&\leq |(\hat{y}_i^1 - \hat{y}_j^1) - (y_i^1 - y_j^1)| + |(\hat{y}_i^0 - \hat{y}_j^0) - (y_i^0 - y_j^0)| \\
&= |(\hat{y}_i^1 - y_i^1) - (\hat{y}_j^1 - y_j^1)| + |(\hat{y}_i^0 - y_i^0) - (\hat{y}_j^0 - y_j^0)| \\
&\leq |\hat{y}_i^1 - y_i^1| + |\hat{y}_j^1 - y_j^1| + |\hat{y}_i^0 - y_i^0| + |\hat{y}_j^0 - y_j^0|.
\end{aligned}
\tag{7}
$$

Meanwhile, another equation variant similar to Eq. 6 is as follows:

$$
\begin{aligned}
&|(\hat{\tau}_i - \hat{\tau}_j) - (\tau_i - \tau_j)| \\
&= |((\hat{y}_i^1 - \hat{y}_j^1) - (\hat{y}_i^0 - \hat{y}_j^0)) - ((y_i^1 - y_j^1) - (y_i^0 - y_j^0))| \\
&= |((\hat{y}_i^1 - y_j^0) - (y_i^1 - \hat{y}_j^0)) - ((\hat{y}_i^0 - y_j^1) - (y_i^0 - \hat{y}_j^1))| \\
&\leq |(\hat{y}_i^1 - y_j^0) - (y_i^1 - \hat{y}_j^0)| + |(\hat{y}_i^0 - y_j^1) - (y_i^0 - \hat{y}_j^1)|.
\end{aligned}
\tag{8}
$$

Following Eq. 8, we can derive:

$$
\begin{aligned}
&|(\hat{\tau}_i - \hat{\tau}_j) - (\tau_i - \tau_j)| \\
&\leq |(\hat{y}_i^1 - y_j^0) - (y_i^1 - \hat{y}_j^0)| + |(\hat{y}_i^0 - y_j^1) - (y_i^0 - \hat{y}_j^1)| \\
&= |(\hat{y}_i^1 - y_i^1) + (\hat{y}_j^0 - y_j^0)| + |(\hat{y}_i^0 - y_i^0) + (\hat{y}_j^1 - y_j^1)| \\
&\leq |\hat{y}_i^1 - y_i^1| + |\hat{y}_j^0 - y_j^0| + |\hat{y}_i^0 - y_i^0| + |\hat{y}_j^1 - y_j^1|.
\end{aligned}
\tag{9}
$$

□

From the Eq. 7 and 9, we can find their last step $|\hat{y}_i^1 - y_i^1| + |\hat{y}_j^1 - y_j^1| + |\hat{y}_i^0 - y_i^0| + |\hat{y}_j^0 - y_j^0|$ are just equivalent to the optimization objective of response MSE loss. This means that conventional MSE loss for response modeling is actually a looser upper error bound than $|(\hat{y}_i^1 - \hat{y}_j^1) - (y_i^1 - y_j^1)| + |(\hat{y}_i^0 - \hat{y}_j^0) - (y_i^0 - y_j^0)|$ and $|(\hat{y}_i^1 - y_j^0) - (y_i^1 - \hat{y}_j^0)| + |(\hat{y}_i^0 - y_j^1) - (y_i^0 - \hat{y}_j^1)|$ from the perspective of uplift ranking, while the latter two are the focus of our following response ranking learning.

4.3 Response Ranking Learning

From the above analysis in Sec. 4.2, we obtain two tighter uplift modeling error bounds from the ranking perspective. Thus, we can transform them into the within-group response ranking loss and cross-group ranking loss between the predicted and the real responses, respectively, to enhance the uplift model's rankability.

4.3.1 Within-group Response Ranking Loss.
We first design the with-group response loss as follows according to Eq. 6:

$$
\mathcal{L}_{wr-rank}^{tt/cc}(i,j) =
\begin{cases}
0, & if(\hat{y}_i^{1/0} - \hat{y}_j^{1/0}) \cdot (y_i^{1/0} - y_j^{1/0}) \geq 0 \\
((\hat{y}_i^{1/0} - \hat{y}_j^{1/0}) - (y_i^{1/0} - y_j^{1/0}))^2, & otherwise.
\end{cases}
\tag{10}
$$

In Eq. 10, for the individual pair (i, j), if the relative order between predicted responses differs from that between true responses, we punish the uplift model with the above alignment loss. Adding the within-group response ranking loss in the treatment group $\mathcal{L}_{wr-rank}^{tt}$ to the within-group response ranking loss in the control group $\mathcal{L}_{wr-rank}^{cc}$, we can obtain the overall $\mathcal{L}_{wr-rank}$:

$$
\mathcal{L}_{wr-rank} = \sum_{i,j \leq |\mathcal{D}^t|} \mathcal{L}_{wr-rank}^{tt}(i,j) + \sum_{i,j \leq |\mathcal{D}^c|} \mathcal{L}_{wr-rank}^{cc}(i,j).
\tag{11}
$$

4.3.2 Cross-group Response Ranking Loss.
Then, we further devise the cross-group response ranking loss as follows according to Eq. 8:

$$
\mathcal{L}_{cr-rank}^{tc/ct}(i,j) =
\begin{cases}
0, & if(\hat{y}_i^{1/0} - y_j^{0/1}) * (y_i^{1/0} - \hat{y}_j^{0/1}) \geq 0 \\
((\hat{y}_i^{1/0} - y_j^{0/1}) - (y_i^{1/0} - \hat{y}_j^{0/1}))^2, & otherwise.
\end{cases}
\tag{12}
$$

Similarly, we punish the uplift model when the predicted responses of individual pair (i, j) cannot meet the cross-group response-related order criteria. Then, adding the treatment-control response ranking loss $\mathcal{L}_{cr-rank}^{tc}$ to the control-treatment response ranking loss $\mathcal{L}_{cr-rank}^{ct}$, we obtain:

$$
\mathcal{L}_{cr-rank} = \sum_{\substack{i \leq |\mathcal{D}^t|, \\ j \leq |\mathcal{D}^c|}} \mathcal{L}_{cr-rank}^{tc}(i,j) + \sum_{\substack{i \leq |\mathcal{D}^c|, \\ j \leq |\mathcal{D}^t|}} \mathcal{L}_{cr-rank}^{ct}(i,j).
\tag{13}
$$

Combining the $\mathcal{L}_{wr-rank}$ and $\mathcal{L}_{cr-rank}$, the final response ranking loss $\mathcal{L}_{r-rank}$ is acquired:

$$
\mathcal{L}_{r-rank} = \mathcal{L}_{wr-rank} + \mathcal{L}_{cr-rank}.
\tag{14}
$$

Note that in the practical implementation of above $\mathcal{L}_{wr-rank}$ and $\mathcal{L}_{cr-rank}$, we randomly sample a certain number of individuals in each group rather than traversing all individuals to accelerate the computation process.

4.4 Uplift Ranking Learning

In this section, we directly dive into the uplift ranking problem and explicitly optimize the uplift model's rankability. It can be easily noted that this is inherently a learning-to-rank problem. The previous solutions mainly include three types: pointwise, pairwise, and listwise. The former two kinds consider the ranking problem from the perspective of a single individual or an individual pair, thus can hardly take all individuals into consideration at a glance. Besides, the pairwise approaches are mainly classification-based and thus have great difficulty distinguishing the inner-pair uplift distance among different pairs. Therefore, they are more suitable for binary responses, i.e., conversion uplift problems. If they are deployed in the revenue uplift scenario where the response is continuous, the performance can hardly be satisfying. The listwise approach, however, can directly minimize the overall ranking error of all individuals in the population and discern the difference in continuous revenue uplift values corresponding to different individuals.

Therefore, we follow the listwise ranking procedure and first model the uplift ranking process as a random process, which means each individual has the probability to be ranked in any position of the whole population list. Here, the ranking probability on the higher position for an individual should be positively related to her uplift score. To tackle this problem, we start from the probability that an individual is ranked at the top-one position of the whole list and propose the loss function to measure the uplift model ranking error. Then, we derive such loss function to an optimizable form with the given data.

4.4.1 Uplift Top One Probability.
Here, referring to [9], we first define the probability that a random individual with the covariate X to be ranked at the top one position:

$$
p_\tau(X) = \frac{e^{s(\tau(X))}}{\sum_X e^{s(\tau(X))}},
\tag{15}
$$

where $s(\cdot)$ is the ranking score function based on the uplift score for X. And it should possess the monotonically increasing property.

4.4.2 Listwise Uplift Ranking Loss.
With the above definition, given the list of true uplift score τ and the list of predicted uplift score $\hat{\tau}$ by the model, we can measure the distance between them as the overall ranking error. Here, we utilize the Cross-Entropy as the measurement and define the listwise uplift ranking loss as follows:

$$\mathcal{L}_{lu-rank} = -\underset{X}{\mathbb{E}}\, p_{\tau}(X)\, ln(p_{\hat{\tau}}(X)). \tag{16}$$

To transform Eq. 16 to an optimizable form, we provide the following derivation process:

Proof.

$$\mathcal{L}_{lu-rank} = -\underset{X}{\mathbb{E}}\, p_{\tau}(X)\, ln(p_{\hat{\tau}}(X))$$
$$= -\underset{X}{\mathbb{E}}\, \frac{e^{s(\tau(X))}}{\sum_X e^{s(\tau(X))}}\, ln(\frac{e^{\hat{s}(\hat{\tau}(X))}}{\sum_X e^{\hat{s}(\hat{\tau}(X))}}). \tag{17}$$

Here, we take $s(\tau(X)) = ln(\tau(X)), \hat{s}(\hat{\tau}(X)) = \hat{\tau}(x)$. Thus, we have:

$$\mathcal{L}_{lu-rank} = -\underset{X}{\mathbb{E}}\, \frac{\tau(X)}{\sum_X \tau(X)}\, ln(\frac{e^{\hat{\tau}(X)}}{\sum_X e^{\hat{\tau}(X)}}). \tag{18}$$

Considering that $\sum_X \tau(X)$ is an unknown constant for each instance of variable X, we remove it from the equation to facilitate the implementation.

$$\mathcal{L}_{lu-rank} = -\underset{X}{\mathbb{E}}\, \tau(X)\, ln(\frac{e^{\hat{\tau}(X)}}{\sum_X e^{\hat{\tau}(X)}})$$
$$= -\underset{X}{\mathbb{E}}\, (Y^1(X) - Y^0(X))\, ln(\frac{e^{\hat{\tau}(X)}}{\sum_X e^{\hat{\tau}(X)}})$$
$$= -(\underset{X}{\mathbb{E}}\, Y^1(X)\, ln(\frac{e^{\hat{\tau}(X)}}{\sum_X e^{\hat{\tau}(X)}}) - \underset{X}{\mathbb{E}}\, Y^0(X)\, ln(\frac{e^{\hat{\tau}(X)}}{\sum_X e^{\hat{\tau}(X)}})). \tag{19}$$

Under the setting of RCT experiments, the assumption that X, X^t, and X^c follow the same distribution holds. Thus, we can obtain:

$$\mathcal{L}_{lu-rank} = -(\underset{X^t}{\mathbb{E}}\, Y^1(X^t)\, ln(\frac{e^{\hat{\tau}(X^t)}}{\sum_X e^{\hat{\tau}(X)}}) - \underset{X^c}{\mathbb{E}}\, Y^0(X^c)\, ln(\frac{e^{\hat{\tau}(X^c)}}{\sum_X e^{\hat{\tau}(X)}}))$$
$$= -(\frac{1}{|\mathcal{D}^t|}\sum_{\mathbf{x}^t \in \mathcal{D}^t} y^1(\mathbf{x}^t)\, ln(\frac{e^{\hat{\tau}(\mathbf{x}^t)}}{\sum_{\mathbf{x} \in \mathcal{D}} e^{\hat{\tau}(\mathbf{x})}})$$
$$- \frac{1}{|\mathcal{D}^c|}\sum_{\mathbf{x}^c \in \mathcal{D}^c} y^0(\mathbf{x}^c)\, ln(\frac{e^{\hat{\tau}(\mathbf{x}^c)}}{\sum_{\mathbf{x} \in \mathcal{D}} e^{\hat{\tau}(\mathbf{x})}})). \tag{20}$$

□

4.5 Overall Training Objective

After obtaining $\mathcal{L}_{ZILN}$ in Sec. 4.1, $\mathcal{L}_{r-rank}$ in Sec. 4.3, and $\mathcal{L}_{lu-rank}$ in Sec. 4.4 sequentially, we add them together and train the parameters $\boldsymbol{\theta}$ of uplift model f with the gradient descent algorithm:

$$\mathcal{L}_{overall} = \mathcal{L}_{ZILN} + \mathcal{L}_{r-rank} + \mathcal{L}_{lu-rank} + \lambda\|\boldsymbol{\theta}\|_2^2, \tag{21}$$

where λ is the coefficient of the L2 regularization term. To better illustrate the process flow of our proposed RERUM framework, we provide the pseudocode in Algorithm 1.

Algorithm 1: RERUM

Input: The observed sample set $\mathcal{D}$ consisting of treatment group $\mathcal{D}^t$ and control group $\mathcal{D}^c$, sample number S

Output: Feature embeddings $\mathbf{e}_f$, representation learning module parameters θ_r, treatment/control response module parameters θ_t, θ_c.

1 **Process**:;
2 Initialize embeddings and parameters $\mathbf{e}_f, \theta_r, \theta_t, \theta_c$;
3 **while** *Not Converged* **do**
4   **for** *batch* $\mathcal{B}^t$ *in* $\mathcal{D}^t$, *batch* $\mathcal{B}^c$ *in* $\mathcal{D}^c$ **do**
5   Obtain predicted (p^1, μ^1, σ^1) and (p^0, μ^0, σ^0) for each sample in $\mathcal{B}^t$ and $\mathcal{B}^c$, respectively;
6   Compute the ZILN regression loss $\mathcal{L}_{ZILN}^t$ and $\mathcal{L}_{ZILN}^c$ for $\mathcal{B}^t$ and $\mathcal{B}^c$ with Eqn. 3 ;
7   Sample S individuals from $\mathcal{B}^t$ and $\mathcal{B}^c$ separately and compute $\mathcal{L}_{wr-rank}^{tt}$ and $\mathcal{L}_{wr-rank}^{cc}$ with Eqn. 10 ;
8   Sample S individuals from $\mathcal{B}^t$ and $\mathcal{B}^c$ separately and compute $\mathcal{L}_{cr-rank}^{tc}$ and $\mathcal{L}_{cr-rank}^{ct}$ with Eqn. 12 ;
9   Utilize $\mathcal{B}^t$ and $\mathcal{B}^c$ to compute the in-batch listwise uplift ranking loss $\mathcal{L}_{lu-rank}$ with Eqn. 20 ;
10   Add above loss terms together to obtain $\mathcal{L}_{overall}$ with Eqn. 21 ;
11   Update parameters $\mathbf{e}_f, \theta_r, \theta_t, \theta_c$ with end-to-end gradient descent on $\mathcal{L}_{overall}$;
12   **end**
13 **end**

---

5 Experiments

In this section, we conduct experiments on three offline datasets and an online fintech marketing platform to show the effectiveness of our method. We mainly focus on the following questions:

- **RQ1:** Can our method outperform different baselines on various ranking-related uplift modeling metrics?
- **RQ2:** How does each proposed module contribute to the overall revenue uplift modeling performance?
- **RQ3:** How do the hyper-parameters influence the performance of our method in different datasets?
- **RQ4:** How do our proposed modules influence the response regression performance?
- **RQ5:** How does our method perform in the online deployment scenario compared with other online service strategies?

5.1 Experiment Setup

5.1.1 Datasets.

- **Hillstrom** [18]: This dataset is derived from an email merchandising campaign that involves 64,000 consumers. The dataset contains three variables describing consumer activity in the following two weeks after the delivery of the email campaign: *visit*, *conversion*, and *spend*. We select the *spend* as the response label which indicates the actual dollar spent in the following two weeks, thus fitting our revenue uplift modeling problem setting. The original dataset has two types of treatment groups *mens_email*,

Table 1: Overall performance of different methods on different uplift-ranking related metrics in various datasets. The best method and best baseline on each metric are marked as bold and underlined, respectively. ↑ refers to the improvement over the best-performing baseline. ∗ indicates the improvements over baselines are statistically significant (*t*-test, *p*-value ≤ 0.05). For ease of illustration, the LIFT@30 for the Product dataset has been divided by 10,000.

Methods	Hillstrom-Men				Hillstrom-Women				Product			
	AUUC	AUQC	KRCC	LIFT@30	AUUC	AUQC	KRCC	LIFT@30	AUUC	AUQC	KRCC	LIFT@30
Causal Forest	0.4492	0.4555	0.0120	0.6376	0.5653	0.5710	0.0776	0.3260	0.6625	0.6809	0.3265	0.9252
S-Learner	0.5301	0.5297	0.0485	0.9123	0.5537	0.5554	0.0654	0.4127	0.6107	0.6088	0.2017	0.8451
T-Learner	0.5565	0.5594	0.0327	1.0417	0.5733	0.5735	0.1018	0.6564	0.6212	0.6374	0.2238	0.8941
TR	0.5484	0.5493	0.0309	1.0388	0.5801	0.5814	0.1317	0.4315	0.6683	0.6983	0.3337	<u>0.9429</u>
TAR	0.5652	0.5659	0.0780	0.7974	0.5739	0.5770	0.0586	0.6366	0.6493	0.6737	0.2986	0.8918
CFR$_{wass}$	0.5661	0.5676	0.0788	0.8721	0.5754	0.5762	0.0606	0.6472	0.6957	0.7018	0.2998	0.8741
CFR$_{mmd}$	0.5760	0.5762	0.0747	0.7351	0.5836	0.5814	0.0788	0.6206	0.6933	0.7025	<u>0.3430</u>	0.8929
StableCFR	0.5772	0.5783	0.0786	0.7524	0.5725	0.5687	0.0965	0.6057	0.6948	<u>0.7067</u>	0.3162	0.8547
CITE	0.5812	0.5793	0.0827	0.7928	0.5785	0.5717	0.1147	0.6742	<u>0.6976</u>	0.6987	0.3348	0.8738
DragonNet	<u>0.6028</u>	<u>0.6042</u>	<u>0.0897</u>	<u>1.3526</u>	<u>0.5858</u>	<u>0.5836</u>	<u>0.1368</u>	<u>0.7123</u>	0.6347	0.6568	0.3253	0.8148
RERUM (TAR)	0.6299*	0.6338*	0.1617*	1.3477*	0.6360*	0.6334*	0.1806*	0.8755*	0.6863*	0.7086*	0.3269*	1.0247*
	↑ 4.50%	↑ 4.90%	↑ 80.27%	↓ 0.36%	↑ 8.57%	↑ 8.53%	↑ 32.02%	↑ 22.91%	↓ 1.62%	↑ 0.27%	↓ 4.69%	↑ 8.68%
RERUM (CFR$_{wass}$)	0.6474*	0.6497*	0.1382*	1.4602*	0.6568*	0.6559*	0.1601*	0.8261*	**0.7563***	**0.7679***	0.3625*	1.1302*
	↑ 7.40%	↑ 7.53%	↑ 54.07%	↑ 7.96%	↑ 12.12%	↑ 12.39%	↑ 17.03%	↑ 15.98%	↑ **8.41%**	↑ **8.66%**	↑ 5.69%	↑ 19.86%
RERUM (CFR$_{mmd}$)	0.6242*	0.6281*	**0.1581***	1.3015*	0.6374*	0.6350*	**0.1692***	0.8252*	0.7554*	0.7647*	0.3588*	**1.2045***
	↑ 3.55%	↑ 3.96%	↑ **76.25%**	↓ 3.78%	↑ 8.81%	↑ 8.81%	↑ 23.68%	↑ 15.85%	↑ 8.29%	↑ 8.21%	↑ 4.61%	↑ **27.74%**
RERUM (DragonNet)	**0.6721***	**0.6753***	0.1434*	**1.5845***	**0.6580***	**0.6566***	0.1664*	**0.9401***	0.7176*	0.7359*	**0.3653***	1.1016*
	↑ **11.50%**	↑ **11.77%**	↑ 59.87%	↑ **17.14%**	↑ **12.33%**	↑ **12.51%**	↑ 21.64%	↑ **31.98%**	↑ 2.87%	↑ 4.13%	↑ **6.50%**	↑ 16.83%

womens_email, and a control group *no_email*. Following [4], we combine the *mens_email* treatment group and *no email* control group as the **Hillstrom-Men** dataset and conduct the similar operation to the *womens_email* treatment group and *no_email* control group to obtain **Hillstrom-Women** dataset.

- **Product**: This is a product dataset that contains over 5,000,000 individuals, collected from the online mutual fund marketing scenario of Tencent FiT. There are more than 1,800 covariate features in the dataset and most of them are categorical. The response label is the amount that an individual pays to purchase funds. The treatment is an incentive coupon.

For these datasets, we perform a 60%/10%/30% split to acquire the training set, validation set, and test set, respectively.

5.1.2 **Baselines**. We compare our RERUM method with the following baselines which are commonly adopted in uplift modeling: Causal Forest [11], S-Learner [22], T-Learner [22], Transformed Response (TR) [3], TAR [29], CFR [29], StableCFR [34], CITE [25], DragonNet [30] . Note that in the following experiments, we mainly take several prestigious deep learning-based uplift models [29, 30] as the backbones of our RERUM, considering their more competitive performance and wider application. We provide the detailed descriptions for these baselines in Appendix B.

5.1.3 **Evaluation Metrics**. In our experiments, we adopt the following four metrics to evaluate the uplift ranking ability of different methods: Area Under the uplift curve (AUUC), Area Under the QINI Curve (AUQC), Kendall Rank Correlation Coefficient (KRCC), and LIFT@h (*h* is set as 30). Following the previous work [13], we take the "jointly, absolute" version of uplift/Qini curves and further formulate the corresponding AUUC and AUQC metrics. For easier computation and fairer comparison, we utilize the *normalized*

AUUC and AUQC approximated over 100 buckets. The detailed evaluation metric introduction is provided in Appendix C.

5.1.4 **Implementation Details**. We conduct each experiment with five different random seeds and take the average performance as the final result. We use the Adam optimizer with the initial learning rate as 0.001. The embedding size for each feature is set as 10, following the conventional design in industrial applications. We implement our method with PyTorch and run it on an NVIDIA Tesla V100 GPU with 32GB memory. Besides, we provide source codes in https://github.com/BokwaiHo/revenue_uplift.

5.2 Results and Analysis

5.2.1 **Overall Performance (RQ1)**. We report the empirical results of our RERUM and baselines on three offline datasets in Table 1. First, we can easily observe that deep learning-based uplift models can reach better performance than tree/meta learner-based models under different settings, which validates the rationality that we take them as the base models of our RERUM. Second, we can observe that, though the uplift modeling capability differences among these base models, our four versions of RERUM with them can all achieve improvement over the best-performing baseline on most metrics and datasets. For example, our RERUM (DragonNet) improves the LIFT@30 metric by 21.98% on average in three datasets. Third, comparing each version of RERUM with its corresponding base model, like RERUM (TAR) and TAR, our RERUM can consistently obtain significant performance gain. Taking the well-known AUUC metric as an example, the RERUM (TAR) brings the 11.45%, 10.82%, 5.70% improvement on Hilstrom-Men, Hillstrom-Women, and Product datasets, respectively. All these results demonstrate the effectiveness and applicability of our RERUM for rankability-enhanced revenue uplift modeling.

Table 2: Ablation study to our three modules under different base models in an incremental manner.

Methods	Hillstrom-Men				Hillstrom-Women				Product			
	AUUC	AUQC	KRCC	LIFT@30	AUUC	AUQC	KRCC	LIFT@30	AUUC	AUQC	KRCC	LIFT@30
TAR	0.5652	0.5659	0.0780	0.7974	0.5739	0.5770	0.0586	0.6366	0.6493	0.6737	0.2986	0.8918
TAR+UR	0.5958	0.5974	0.1083	0.9098	0.5980	0.5961	0.0947	0.7067	0.6725	0.7028	0.3091	1.0485
TAR+UR+RR	0.6187	0.6199	0.1527	1.1813	0.6204	0.6197	0.1390	0.7668	0.6780	0.7071	0.3172	1.0610
RERUM (TAR)	0.6299	0.6338	0.1617	1.3477	0.6360	0.6334	0.1806	0.8755	0.6863	0.7086	0.3269	1.0247
CFR_{wass}	0.5661	0.5676	0.0788	0.8721	0.5754	0.5762	0.0606	0.6472	0.6957	0.7018	0.2998	0.8741
CFR_{wass}+UR	0.6027	0.6042	0.0844	0.9020	0.5973	0.5957	0.1247	0.7261	0.7150	0.7301	0.3451	0.9721
CFR_{wass}+UR+RR	0.6150	0.6191	0.1349	1.1506	0.6144	0.6132	0.1490	0.7442	0.7384	0.7459	0.3507	1.0291
RERUM (CFR_{wass})	0.6474	0.6497	0.1382	1.4602	0.6568	0.6559	0.1601	0.8261	0.7563	0.7679	0.3625	1.1302
CFR_{mmd}	0.5760	0.5762	0.0747	0.7351	0.5836	0.5814	0.0788	0.6206	0.6933	0.7025	0.3430	0.8929
CFR_{mmd}+UR	0.5981	0.5998	0.1160	0.9163	0.6081	0.5998	0.1301	0.7193	0.7127	0.7212	0.3697	1.0126
CFR_{mmd}+UR+RR	0.6194	0.6206	0.1568	1.1791	0.6137	0.6114	0.1525	0.7547	0.7285	0.7357	0.3014	1.0166
RERUM (CFR_{mmd})	0.6242	0.6281	0.1581	1.3015	0.6374	0.6350	0.1692	0.8252	0.7554	0.7647	0.3588	1.2045
DragonNet	0.6028	0.6042	0.0897	1.3526	0.5858	0.5836	0.1368	0.7123	0.6347	0.6568	0.3253	0.8148
DragonNet+UR	0.6534	0.6555	0.1155	1.4343	0.6184	0.6181	0.1695	0.7891	0.6623	0.6880	0.3224	0.9277
DragonNet+UR+RR	0.6684	0.6713	0.1269	1.5656	0.6401	0.6394	0.1574	0.9017	0.6805	0.7040	0.3359	1.0444
RERUM (DragonNet)	0.6721	0.6753	0.1434	1.5845	0.6580	0.6566	0.1664	0.9401	0.7176	0.7359	0.3653	1.1016

5.2.2 **Ablation Study (RQ2)**. We conduct the ablation study by incrementally adding the uplift ranking learning module (UR), response ranking learning module (RR), and ZILN loss along with its corresponding response modeling framework, to the different base models in a sequential manner. The experimental results on three datasets are reported in Table 2. First, we can find that after the introduction of each module, the performance can all be strengthened to some extent on four ranking-related metrics, which demonstrates that our three contributions can all benefit the revenue uplift modeling. Second, comparing the performance of X+UR+RR and RERUM (X) (X indicates the base model), we can notice that the ZILN response regression module can effectively enhance the rankability of the uplift model, though it does not consider the uplift ranking explicitly. This also validates our motivation in Sec. 4.1, that is, accurate revenue response modeling plays a crucial role in uplift modeling.

5.2.3 **Hyperparameter Sensitivity Analysis (RQ3)**. We conduct the sensitivity analysis on two essential hyperparameters: batch size B which influences the estimation of $\mathcal{L}_{lu-rank}$ in Sec. 4.4 and the number of samples S in Sec. 4.3. Due to the space limitation, we only present the results of RERUM (CFR_{mmd}) and RERUM (DragonNet) regarding AUUC and LIFT@30 metrics. The results are illustrated in Fig. 3. First, we can observe that both RERUM (CFR_{mmd}) and RERUM (DragonNet) are relatively robust to the hyperparameter selection, no matter of AUUC or LIFT@30 metric. Second, it can be noticed that under different hyperparameters, the performance of RERUM (DragonNet) fluctuates a little bit more than RERUM (CFR_{mmd}). This is due to the target regularization in DragonNet itself, which is also consistent with previous research.

5.2.4 **Response Regression Performance Analysis (RQ4)**. As emphasized in the Sec. 1, the focus of this work is enhancing the uplift model's rankability and all three proposed modules are surrounding how to boost the model's capability to identify the most
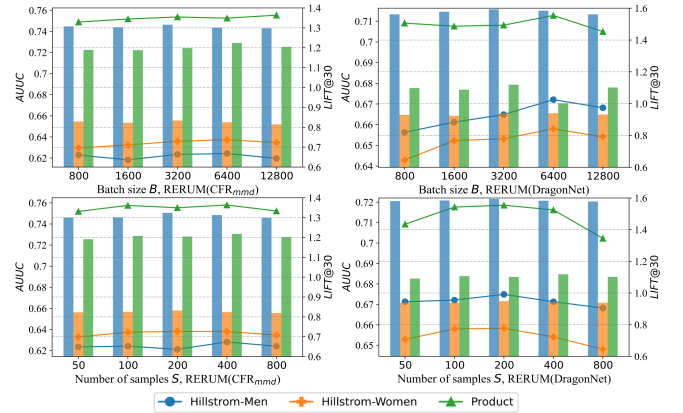

Figure 3: The hyperparameter sensitivity analysis results. The line chart corresponds to the AUUC metric and the column chart corresponds to the LIFT@30 metric.

susceptible individuals to the treatment. Even utilizing ZILN instead of the conventional MSE to conduct the regression is also for improving ranking performance via accurately predicting the revenue response. Even though, investigating how such modules (especially UR and RR modules which are not specially designed for response regression) fare on the underlying response regression task can help provide the more comprehensive understanding to our methods. Therefore, we compare the performance of X (X indicates the base model), X+ZILN, X+ZILN+UR, RERUM(X), on the commonly used regression metric Mean Absolute Percentage Error (MAPE) of the response prediction (lower is better). Note that all base models X utilize the conventional MSE as the response regression loss according to their original papers. The detailed experimental results are provided in Table 3. From the Table 3, we can find that X with ZILN as the regression loss can significantly outperform the X with the MSE as the regression loss on the MAPE

Table 3: Impacts of our proposed modules to response regression accuracy under different base models. Mean Absolute Percentage Error (MAPE) is taken as the metric.

Methods	Hillstrom-Men	Hillstrom-Women	Product
TAR	0.2142	0.2087	0.4429
TAR+ZILN	0.0586	0.0514	0.1378
TAR+ZILN+UR	0.0614	0.0575	0.1476
TAR+ZILN+RR	0.0608	0.0550	0.1423
RERUM (TAR)	0.0667	0.0634	0.1535
CFR_{wass}	0.2357	0.2024	0.3966
CFR_{wass}+ZILN	0.0651	0.0488	0.1102
CFR_{wass}+ZILN+UR	0.0685	0.0524	0.1197
CFR_{wass}+ZILN+RR	0.0677	0.0506	0.1174
RERUM (CFR_{wass})	0.0741	0.0589	0.1265
CFR_{mmd}	0.2456	0.2120	0.4251
CFR_{mmd}+ZILN	0.0712	0.0634	0.1206
CFR_{mmd}+ZILN+UR	0.0785	0.0697	0.1385
CFR_{mmd}+ZILN+RR	0.0764	0.0675	0.1310
RERUM (CFR_{mmd})	0.0823	0.0731	0.1436
DragonNet	0.1955	0.2037	0.3754
DragonNet+ZILN	0.0483	0.0528	0.0986
DragonNet+ZILN+UR	0.0545	0.0617	0.1028
DragonNet+ZILN+RR	0.0516	0.0574	0.1001
RERUM (DragonNet)	0.0628	0.0654	0.1109

metric. This demonstrates that the ZILN loss can effectively help predict the revenue responses more accurately, thus finally benefiting the uplift ranking. Besides, the introduction of UR and RR will influence the performance on MAPE metric a little bit. Because their main goal is to increase the uplift ranking accuracy instead of the response regression. As long as they can facilitate the uplift ranking performance (shown in the Table 1 of the paper), a small degree of decrease in response regression is still tolerable. Even though, the regression performance (MAPE) of X+ZILN+UR, X+ZILN+RR, RERUM(X) is still better than X with MSE as the regression loss.

5.2.5 ***Online Deployment (RQ5).*** This is a mutual fund sales scenario with the notification redpoint as the treatment in the wealth management business of our deployment platform, Tencent FiT, one of the world's largest online fintech marketing platforms. Here, the response variable is the sales revenue. We first identify the top 2% ranked individuals from the whole population of around 400 million users by the uplift model. Then, we randomly split such people into the treatment group and control group. The difference between the average response in such two groups in the following certain length of period (1 month), also known as the sales revenue LIFT@2, reflects the rankability of the model. The online deployment platform is illustrated in Fig. 4. To enhance the reliability and validity, we conduct three times of such marketing campaigns to demonstrate the effectiveness of our RERUM.

Online Experimental Results Analysis. From the results shown in Table 4, we can observe that our RERUM achieves consistent improvement over the state-of-the-art (SOTA) online base model
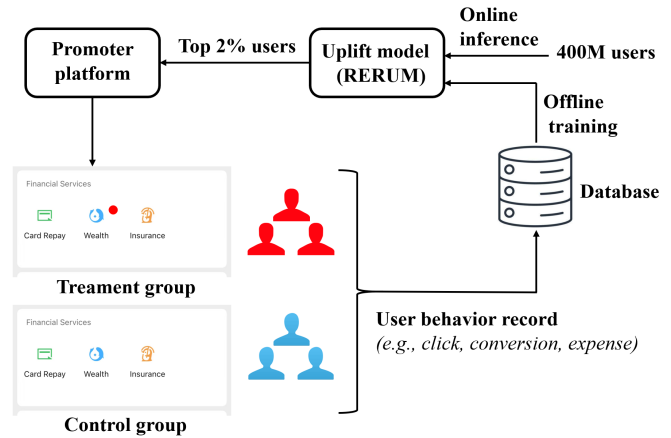
Figure 4: Overview of the online deployment platform.

Table 4: Performance of SOTA online base model and our RERUM on sales revenue LIFT@2 in online deployment.

Methods	Campaign 1	Campaign 2	Campaign 3
Base	24566	30137	21971
RERUM(Base) Improvement	26826 ↑ 9.20%	41360 ↑ 37.24%	25361 ↑ 15.43%

across three campaigns. The average improvement reaches 20.61% on the sales revenue LIFT@2 metric. Besides, the RERUM brings 430 million dollar assets under management (AUM) gain each month. Such empirical results strongly demonstrate the effectiveness of our RERUM in real-world applications.

6 Conclusions and Future Work

Revenue uplift modeling has been recognized as of great importance for online marketing. Especially, the uplift ranking within the whole population is the core to identifying the sensitive individuals to interventions. However, previous methods suffer from the low revenue response prediction precision and the neglection to the uplift ranking accuracy among individuals. In this paper, we have three contributions to address this problem: 1) modeling the response regression with a ZILN loss and a customized network framework that can be adapted with different base uplift models; 2) analyzing the uplift ranking errors and using two tighter response ranking losses to augment the vanilla regression term; 3) directly modeling the listwise uplift ranking among the whole population. The extensive experiments on both offline datasets and one of the world's largest online fintech marketing platforms demonstrate the effectiveness of our proposed method. In the future, we prepare to explore the revenue uplift modeling with multiple treatments and its application in cross-domain scenarios.

Acknowledgments

References

[1] Hervé Abdi. 2007. The Kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA* (2007), 508–510.

[2] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. 2015. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* 33, 4 (2015), 485–505.

[3] Susan Athey and Guido W Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. *stat* 1050, 5 (2015), 1–26.

[4] 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 & Data Mining*. 55–65.

[5] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22, 14 (2006), e49–e57.

[6] Christopher Burges, Robert Ragno, and Quoc Le. 2006. Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems* 19 (2006).

[7] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*. 89–96.

[8] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.

[9] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*. 129–136.

[10] Ziqiang Cui, Xing Tang, Yang Qiao, Bowei He, Liang Chen, Xiuqiang He, and Chen Ma. 2024. Treatment-Aware Hyperbolic Representation Learning for Causal Effect Estimation with Social Networks. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 289–297.

[11] Jonathan MV Davis and Sara B Heller. 2017. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review* 107, 5 (2017), 546–550.

[12] Angus Deaton and Nancy Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social science & medicine* 210 (2018), 2–21.

[13] 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.

[14] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research* 4, Nov (2003), 933–969.

[15] Robin Marco Gubela and Stefan Lessmann. 2020. Interpretable Multiple Treatment Revenue Uplift Modeling. In *26th Americas Conference on Information Systems, AMCIS 2020, Virtual Conference, August 15-17, 2020*, Bonnie Brinton Anderson, Jason Thatcher, Rayman D. Meservy, Kathy Chudoba, Kelly J. Fadel, and Sue Brown (Eds.). Association for Information Systems. https://aisel.aisnet.org/amcis2020/data_science_analytics_for_decision_support/data_science_analytics_for_decision_support/18

[16] Robin M Gubela, Stefan Lessmann, and Szymon Jaroszewicz. 2020. Response transformation and profit decomposition for revenue uplift modeling. *European Journal of Operational Research* 283, 2 (2020), 647–661.

[17] Pierre Gutierrez and Jean-Yves Gérardy. 2017. Causal inference and uplift modelling: A review of the literature. In *International conference on predictive applications and APIs*. PMLR, 1–13.

[18] Kevin Hillstrom. 2008. The MineThatData e-mail analytics and data mining challenge. *MineThatData blog* (2008).

[19] Ziniu Hu, Yang Wang, Qu Peng, and Hang Li. 2019. Unbiased lambdamart: an unbiased pairwise learning-to-rank algorithm. In *The World Wide Web Conference*. 2830–2836.

[20] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 133–142.

[21] Kathleen Kane, Victor SY Lo, and Jane Zheng. 2014. Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics* 2 (2014), 218–238.

[22] 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.

[23] Yanyan Lan, Yadong Zhu, Jiafeng Guo, Shuzi Niu, and Xueqi Cheng. 2014. Position-Aware ListMLE: A Sequential Learning Process for Ranking.. In *UAI*, Vol. 14. 449–458.

[24] Ping Li, Qiang Wu, and Christopher Burges. 2007. Mcrank: Learning to rank using multiple classification and gradient boosting. *Advances in neural information processing systems* 20 (2007).

[25] Xinshu Li and Lina Yao. 2022. Contrastive individual treatment effects estimation. In *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1053–1058.

[26] Jing Ma, Mengting Wan, Longqi Yang, Jundong Li, Brent Hecht, and Jaime Teevan. 2022. Learning causal effects on hypergraphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1202–1212.

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

[28] Piotr Rzepakowski and Szymon Jaroszewicz. 2010. Decision trees for uplift modeling. In *2010 IEEE International Conference on Data Mining*. IEEE, 441–450.

[29] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.

[30] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems* 32 (2019).

[31] Bonnie Sibbald and Martin Roland. 1998. Understanding controlled trials. Why are randomised controlled trials important? *BMJ: British Medical Journal* 316, 7126 (1998), 201.

[32] SS Vallender. 1974. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications* 18, 4 (1974), 784–786.

[33] Xiaojing Wang, Tianqi Liu, and Jingang Miao. 2019. A deep probabilistic model for customer lifetime value prediction. *arXiv preprint arXiv:1912.07753* (2019).

[34] Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. 2023. Stable Estimation of Heterogeneous Treatment Effects. (2023).

[35] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*. 1192–1199.

[36] Jun Xu and Hang Li. 2007. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 391–398.

[37] Hao Zhou, Shaoming Li, Guibin Jiang, Jiaqi Zheng, and Dong Wang. 2023. Direct Heterogeneous Causal Learning for Resource Allocation Problems in Marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5446–5454.

A Main Notations

To facilitate comprehension, we summarize the main notations used in this paper in Table 5.

Table 5: Major notations.

f	The uplift model
τ	The true uplift based on the potential responses
$\mathcal{D}^t, \mathcal{D}^c$	The treatment group and control group
$\mathbf{x}^t, X^t$	The covariate vectors of users in treatment group and their corresponding random variable
$\mathbf{x}^c, X^c$	The covariate vectors of users in control group and their corresponding random variable
y^1, Y^1	The potential response instances under treatment and their corresponding random variable
y^0, Y^0	The potential response instances under control and their corresponding random variable
Y, X, T	The general random variables that indicate the response, covariate, and treatment, respectively
Y_i, X_i, T_i	The random variables that indicate the individual i's response, covariate, and treatment, respectively

B Descriptions to Baselines

We provide the detailed descriptions to all baselines as follows:

- **Causal Forest** [11]: Causal Forest is a non-parametric Random Forest-based tree model that directly estimates the treatment effect, which is one of the most representative tree-based uplift models in many areas like economics and social science.

- **S-Learner** [22]: S-Learner is a kind of meta-learner method that uses a single estimator to estimate the response without giving the treatment a special role. The uplift is estimated by the difference between changed treatments with fixed covariates.

- **T-Learner** [22]: T-Learner is another type of meta-learner method. Different from S-Learner, T-Learner uses two estimators for the treatment group and control group, respectively.
- **Transformed Response (TR)** [3]: TR is a special technique that transforms the observed response Y to Y^*, such that the uplift equals the conditional expectation of the transformed response from the expectation perspective.
- **TAR** [29]: TAR is a commonly used deep learning-based uplift model. Compared with T-Learner, it omits the additional imputed treatment effects fitting sub-models but introduces the shared layers for treated and control response networks. The shared network parameters could help alleviate the sample imbalance issue between the treatment and control groups.
- **CFR** [29]: On the basis of TAR, CFR applies an additional loss to TAR, which forces the learned treated and control covariate distributions to be closer. We report the CFR performance using two distribution distance measurement loss functions, Wasserstein [32] (denoted as CFR_{wass}) and Maximum Mean Discrepancy (MMD) [5] (denoted as CFR_{mmd}).
- **StableCFR** [34]: StableCFR upsamples the underrepresented data with uniform sampling and balances covariates by using an epsilon-greedy matching mechanism to achieve higher uplift effect estimation accuracy for the underrepresented population.
- **CITE** [25]: CITE is based on the contrastive task designed for causal inference, it fully exploits the self-supervision information hidden in data to achieve balanced and predictive representations while appropriately leveraging causal prior knowledge.
- **DragonNet** [30]: Dragonnet exploits the sufficiency of the propensity score for estimation adjustment, and uses a regularization procedure based on non-parametric estimation theory which can guarantee desirable asymptotic properties.

C Descriptions to Evaluation Metrics

We provide the detailed descriptions to our four uplift ranking evaluation metrics as follows:

- **Area Under the Uplift Curve (AUUC)**: In uplift research, to evaluate the rankability of the uplift model f, one can first plot an uplift curve which ranks individual samples descendingly according to *predicted* uplift $\hat{\tau}$ (in X-axis) and cumulatively sums the *observed* uplift (in Y-axis) [4]. The AUUC is then the area under this curve. There are actually multiple variants of uplift curves proposed in the recent literature. Their differences mainly lie in 1) if ranking the data separately per group or jointly over all data, and 2) if expressing volumes in absolute or relative numbers [13]. In this work, we take the "jointly, absolute" uplift curves in [13] and formulate the AUUC. First, we denote the total number of treated and control instances, among the top-k individuals $\pi(\mathcal{D}, k)$ ranked by uplift model f over the whole dataset $\mathcal{D}$ as

$$
\begin{aligned}
N_\pi^T(\mathcal{D}, k) &= \sum_{(\mathbf{x}_i, t_i, y_i) \in \pi(\mathcal{D}, k)} \mathbb{I}(t_i = 1), \\
N_\pi^C(\mathcal{D}, k) &= \sum_{(\mathbf{x}_i, t_i, y_i) \in \pi(\mathcal{D}, k)} \mathbb{I}(t_i = 0),
\end{aligned}
\tag{22}
$$

where the $\mathbb{I}(\cdot)$ is the indicator function. Then, the response summation of treated and control individuals in $\pi(\mathcal{D}, k)$ are:

$$
\begin{aligned}
R_\pi^T(\mathcal{D}, k) &= \sum_{(\mathbf{x}_i, t_i, y_i) \in \pi(\mathcal{D}, k)} y_i \mathbb{I}(t_i = 1), \\
R_\pi^C(\mathcal{D}, k) &= \sum_{(\mathbf{x}_i, t_i, y_i) \in \pi(\mathcal{D}, k)} y_i \mathbb{I}(t_i = 0),
\end{aligned}
\tag{23}
$$

where y_i is the observed response. Thus, the values for the uplift curve can be obtained with:

$$
V_u(f, k) = \left(\frac{R_\pi^T(\mathcal{D}, k)}{N_\pi^T(\mathcal{D}, k)} - \frac{R_\pi^C(\mathcal{D}, k)}{N_\pi^C(\mathcal{D}, k)} \right) * (N_\pi^T(\mathcal{D}, k) + N_\pi^C(\mathcal{D}, k)).
\tag{24}
$$

Furthermore, the AUUC is formulated with:

$$
AUUC(f, k) = \int_0^1 V_u(f, x) dx = \sum_{k=1}^n V_u(f, k) \approx \sum_{p=1}^{100} V_u\left(f, \frac{p}{100}\right).
\tag{25}
$$

For the separate setting, we can make the above approximation over 100 buckets to estimate the AUUC. In this paper, we utilize the *normalized* AUUC for the fairer comparison.

- **Area Under the QINI Curve (AUQC)**: Similar to the above AUUC, we follow [13] and take the "jointly, absolute" Qini curve to formulate AUQC metric. Based on the previous definition in Eq. 22 and 23, we define the values for the Qini curve as:

$$
V_q(f, k) = R_\pi^T(\mathcal{D}, k) - R_\pi^C(\mathcal{D}, k) * \frac{N_\pi^T(\mathcal{D}, k)}{N_\pi^C(\mathcal{D}, k)}.
\tag{26}
$$

Thus, the AUQC can be obtained as:

$$
AUQC(f, k) = \int_0^1 V_q(f, x) dx = \sum_{k=1}^n V_q(f, k) \approx \sum_{p=1}^{100} V_q\left(f, \frac{p}{100}\right).
\tag{27}
$$

We use the *normalized* AUQC to compare different methods.

- **Kendall Rank Correlation Coefficient (KRCC)**: The KRCC [1] measures the similarity between the rank by predicted uplift scores and the rank by the approximated true uplift scores for all individuals. To facilitate the computation, we split the whole population into 100 buckets and use the treatment and control group data in each bucket to approximate the true uplift effect.
- **LIFT@h**: This metric measures the difference between the mean response of treated individuals and that of controlled individuals in the top h percentile of all individuals ranked by the uplift model. It has been widely employed in many industrial scenarios, because it can explicitly reflect the model's rankability, especially for a certain proportion of targeted people. Here, we take h as 30.

D Supplementary Ablation Study

In addition to the ablation study conducted in Sec. 5.2.2, we also adjust the module stacking order to further highlight the effect of each module to the revenue uplift modeling. In detail, we incrementally add the ZILN regression module (ZILN), uplift ranking learning module (UR), and response ranking learning module (RR) to the vanilla base model X. The corresponding experimental results are illustrated in Table 6. First, it is evident that X+ZILN can achieve better performance than vanilla X under almost all ranking-related metrics and datasets, which implies that the ZILN module

Table 6: Ablation study to our three modules under different base models by incrementally adding ZILN regression module (ZILN), uplift ranking learning module (UR), and response ranking learning module (RR).

Methods	Hillstrom-Men				Hillstrom-Women				Product			
	AUUC	AUQC	KRCC	LIFT@30	AUUC	AUQC	KRCC	LIFT@30	AUUC	AUQC	KRCC	LIFT@30
TAR	0.5652	0.5659	0.0780	0.7974	0.5739	0.5770	0.0586	0.6366	0.6493	0.6737	0.2986	0.8918
TAR+ZILN	0.5741	0.5784	0.0912	0.8216	0.5972	0.5946	0.0831	0.6687	0.6679	0.6910	0.3105	0.9211
TAR+ZILN+UR	0.6028	0.6131	0.1407	1.1625	0.6206	0.6218	0.1577	0.7384	0.6813	0.7022	0.3209	0.9885
RERUM (TAR)	0.6299	0.6338	0.1617	1.3477	0.6360	0.6334	0.1806	0.8755	0.6863	0.7086	0.3269	1.0247
CFR_{wass}	0.5661	0.5676	0.0788	0.8721	0.5754	0.5762	0.0606	0.6472	0.6957	0.7018	0.2998	0.8741
CFR_{wass}+ZILN	0.5816	0.5924	0.1052	0.9002	0.5904	0.5933	0.0782	0.6671	0.7025	0.7068	0.3112	0.8946
CFR_{wass}+ZILN+UR	0.6214	0.6282	0.1237	1.2841	0.6335	0.6361	0.1323	0.7835	0.7364	0.7438	0.3442	1.0824
RERUM (CFR_{wass})	0.6474	0.6497	0.1382	1.4602	0.6568	0.6559	0.1601	0.8261	0.7563	0.7679	0.3625	1.1302
CFR_{mmd}	0.5760	0.5762	0.0747	0.7351	0.5836	0.5814	0.0788	0.6206	0.6933	0.7025	0.3430	0.8929
CFR_{mmd}+ZILN	0.5942	0.5964	0.1138	0.7590	0.5933	0.5928	0.0923	0.6418	0.7023	0.7186	0.3491	0.9112
CFR_{mmd}+ZILN+UR	0.6125	0.6153	0.1406	1.1247	0.6213	0.6204	0.1429	0.7846	0.7343	0.7425	0.3527	1.1136
RERUM (CFR_{mmd})	0.6242	0.6281	0.1581	1.3015	0.6374	0.6350	0.1692	0.8252	0.7554	0.7647	0.3588	1.2045
DragonNet	0.6028	0.6042	0.0897	1.3526	0.5858	0.5836	0.1368	0.7123	0.6347	0.6568	0.3253	0.8148
DragonNet+ZILN	0.6204	0.6227	0.1235	1.4167	0.6124	0.6140	0.1435	0.7482	0.6627	0.6835	0.3349	0.8582
DragonNet+ZILN+UR	0.6587	0.6603	0.1395	1.5274	0.6402	0.6419	0.1581	0.8826	0.6935	0.7148	0.3564	1.0274
RERUM (DragonNet)	0.6721	0.6753	0.1434	1.5845	0.6580	0.6566	0.1664	0.9401	0.7176	0.7359	0.3653	1.1016

Table 7: Ablation study to within-group response ranking loss and cross-group response ranking loss, respectively.

Methods	Hillstrom-Men				Hillstrom-Women				Product			
	AUUC	AUQC	KRCC	LIFT@30	AUUC	AUQC	KRCC	LIFT@30	AUUC	AUQC	KRCC	LIFT@30
RERUM (TAR) w/o wr-rank	0.6203	0.6241	0.1293	1.1927	0.6214	0.6207	0.1561	0.8382	0.6825	0.7030	0.3184	1.0032
RERUM (TAR) w/o cr-rank	0.6228	0.6272	0.1385	1.2364	0.6256	0.6243	0.1620	0.8541	0.6839	0.7051	0.3207	1.0125
RERUM (TAR)	0.6299	0.6338	0.1617	1.3477	0.6360	0.6334	0.1806	0.8755	0.6863	0.7086	0.3269	1.0247
RERUM (CFR_{wass}) w/o wr-rank	0.6384	0.6401	0.1185	1.3425	0.6437	0.6430	0.1477	0.8133	0.7489	0.7576	0.3539	1.0986
RERUM (CFR_{wass}) w/o cr-rank	0.6405	0.6422	0.1236	1.3861	0.6482	0.6489	0.1542	0.8192	0.7521	0.7605	0.3574	1.1133
RERUM (CFR_{wass})	0.6474	0.6497	0.1382	1.4602	0.6568	0.6559	0.1601	0.8261	0.7563	0.7679	0.3625	1.1302
RERUM (CFR_{mmd}) w/o wr-rank	0.6188	0.6219	0.1395	1.1721	0.6288	0.6274	0.1583	0.8160	0.7448	0.7539	0.3487	1.1208
RERUM (CFR_{mmd}) w/o cr-rank	0.6215	0.6240	0.1467	1.2258	0.6313	0.6299	0.1616	0.8192	0.7483	0.7565	0.3512	1.1519
RERUM (CFR_{mmd})	0.6242	0.6281	0.1581	1.3015	0.6374	0.6350	0.1692	0.8252	0.7554	0.7647	0.3588	1.2045
RERUM (DragonNet) w/o wr-rank	0.6639	0.6672	0.1298	1.4763	0.6465	0.6482	0.1545	0.9287	0.7052	0.7274	0.3561	0.9852
RERUM (DragonNet) w/o cr-rank	0.6673	0.6701	0.1365	1.5272	0.6519	0.6528	0.1617	0.9340	0.7089	0.7302	0.3596	1.0324
RERUM (DragonNet)	0.6721	0.6753	0.1434	1.5845	0.6580	0.6566	0.1664	0.9401	0.7176	0.7359	0.3653	1.1016

can indeed improve the rankability of uplift models. This is actually consistent with our discussion in Sec. 4.1. Second, we can also notice that integrating each module into the overall framework can respectively bring about improvement in ranking performance, which further verifies their effectiveness.

To demonstrate that both within-group response ranking loss (wr-rank) and cross-group ranking loss (cr-rank) in Sec. 4.3 contribute positively to the final uplift ranking performance, we specifically remove them from the overall scheme respectively and observe the corresponding results. The complete version of the results is present in Table 7. From this table, we can first observe that both RERUM(X) w/o wr-rank and RERUM(X) w/o cr-rank achieve lower ranking performance than RERUM(X), which demonstrates that both wr-rank and cr-rank contribute positively to the final performance, i.e., both wr-rank and cr-rank are necessary. Besides, we

can find that the performance drop of removing wr-rank is a little more than that brought by removing cr-ranking, which implies that the effect of cr-rank is a little bit weaker than wr-rank.