Robustness-enhanced Uplift Modeling with Adversarial Feature Desensitization

Zexu Sun, Bowei He, Ming Ma, Jiakai Tang, Yuchen Wang, Chen Ma, Dugang Liu

Gaoling School of Artificial Intelligence, Renmin University of China
Department of Computer Science, City University of Hong Kong
Kuaishou Technology

*Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University

Abstract—Uplift modeling has shown very promising results in online marketing. However, most existing works are prone to the robustness challenge in some practical applications. In this paper, we first present a possible explanation for the above phenomenon. We verify that there is a feature sensitivity problem in online marketing using different real-world datasets, where the perturbation of some key features will seriously cause the performance of the uplift model and even cause the opposite trend. To solve the above problem, we propose a novel robustness-enhanced uplift modeling framework with adversarial feature desensitization (RUAD). Specifically, our RUAD can more effectively alleviate the feature sensitivity of the uplift model through two customized modules, including a feature selection module with joint multi-label modeling to identify a key subset from the input features and an adversarial feature desensitization module using adversarial training and soft interpolation operations to enhance the robustness of the model against this selected subset of features. Finally, we conduct extensive experiments on a public dataset and a real product dataset to verify the effectiveness of our RUAD in online marketing. In addition, we also demonstrate the robustness of our RUAD to the feature sensitivity, as well as the compatibility with different uplift models.

Index Terms—Uplift modeling, Robustness, Adversarial training, Feature desensitization

I. INTRODUCTION

One of the critical tasks in each service platform is to increase user engagement and platform revenue through online marketing, which uses some well-designed incentives and then delivers them to the platform users, such as coupons, discounts, and bonuses [1]. Since each incentive usually comes with a cost, successful online marketing must accurately find the corresponding sensitive user group for each to avoid ineffective delivery. To achieve this goal, an important step is that the marketing model needs to identify the change in the user's response caused by different incentives, and only deliver each incentive to its high-uplift users. This involves a typical causal inference problem, i.e., the estimation of the individual treatment effect (ITE) (also known as the uplift), since we usually only observe one type of user response in practice, which may be for a certain incentive (i.e., treatment group) or no incentive (i.e., control group). Therefore, previous works have proposed uplift modeling and verified its effectiveness in online marketing [2].

The existing methods for uplift modeling can broadly be categorized into three research lines: 1) Meta-learner-based. The basic idea of this line is to estimate the users’ responses by using existing predictive models as the base learner. Two of the most representative methods are S-Learner and T-Learner [3], which adopt a global base learner and two base learners corresponding to the treatment and control groups, respectively. 2) Tree-based. The basic idea of this line is to employ a hierarchical tree structure to systematically partition the user population into sub-populations that exhibit sensitivity to specific treatments [4]. An essential step involves modeling the uplift directly by applying diverse splitting criteria. 3) Neural network-based. The basic idea of this line is to leverage the power of neural networks to develop estimators that are both intricate and versatile in predicting the user's response. Note that most of them can be seen as the variants of meta-learners. We focus on the neural network-based line because they can be more flexibly adapted to modeling the complex feature interactions in many industrial systems. Furthermore, due to the widespread use of various neural network models in these systems, research on this line is also easier to seamlessly integrate than alternative lines.

Although existing works on uplift modeling have shown very promising results, they generally suffer from the robustness challenge in many real-world scenarios [5], and little research has been conducted to reveal how such challenges arise. In this paper, we first identify a feature sensitivity problem in the uplift model as a possible explanation for the above phenomenon using different real-world datasets. Specifically, for each dataset, we randomly select 30% of all continuous-valued features and apply a Gaussian noise with \( \eta \sim \mathcal{N}(0, 0.05^2) \) as the perturbation to them. We repeat this process multiple times to obtain a set of copies with different feature subsets. Finally, we train the same uplift models for each copy and compare their performance with that obtained on the original dataset. Due to space limitations, we show the results of using S-Learner as the uplift model on the Production dataset used in the experiment in Fig. 1, and similar
results are also found on other datasets or uplift models. We can find that there are some sensitive key features and a slight perturbation to them will seriously affect the performance of the uplift model, and even an opposite trend appears.

The above empirical findings suggest that the sensitivity of uplift models to these key features may be one of the important reasons for their robustness challenges. Therefore, to alleviate the feature sensitivity problem, we propose a novel robustness-enhanced uplift modeling framework with adversarial feature desensitization, or RUAD for short. Our RUAD contains two new custom modules that match our empirical findings and can be integrated with most existing uplift models to improve their robustness. Specifically, a feature selection module with joint multi-label training will be used to identify the desired set of key sensitive features from the original dataset under the supervision of a trade-off optimization objective. Then, an adversarial feature desensitization module performs an adversarial training operation and a soft interpolation operation based on the selected subset of features, to force the model to reduce sensitivity to them, thus effectively truncating a key source of robustness challenges. Finally, we experimentally verify the effectiveness of our RUAD on a public dataset and a real product dataset.

II. PRELIMINARIES

To formalize the problem, we follow the Neyman-Rubin potential outcome framework [6], to define the uplift modeling problem. Let the observed sample set be \( D = \{(x_i, t_i, y_i)\}_{i=1}^n \). Without loss of generality, for each sample, assuming \( y_i \in Y \subset \mathbb{R} \) is a continuous response variable, \( x_i \in X \subset \mathbb{R}^N \) is a vector of features, and \( t_i \in \{0, 1\} \) denotes the treatment indicator variable, i.e., whether to get an incentive delivery. Note that the proposed framework can also be easily extended to other types of uplift modeling problems. For a user \( i \), the change in user response caused by an incentive \( t_i \), i.e., individual treatment effect or uplift, denoted as \( \tau_i \), is defined as the difference between the treatment response and the control response,

\[
\tau_i = y_i(1) - y_i(0),
\]

where \( y_i(0) \) and \( y_i(1) \) are the user responses of the control and treatment groups, respectively.

In the ideal world, i.e., obtaining the responses of a user in both groups simultaneously, we can easily determine the uplift \( \tau_i \) based on Eq.(1). However, in the real world, usually, only one of the two responses is observed for any one user. For example, if we have observed the response of a customer who receives the discount, it is impossible for us to observe the response of the same customer when they do not receive a discount, where such responses are often referred to as counterfactual responses. Therefore, the observed response can also be described as,

\[
y_i = t_i y_i(1) + (1 - t_i) y_i(0). \tag{2}
\]

For the brevity of notation, we will omit the subscript \( i \) in the following if no ambiguity arises.

As mentioned above, the uplift \( \tau \) is not identifiable since the observed response \( y \) is only one of the two necessary terms (i.e., \( y(1) \) and \( y(0) \)). Fortunately, with some appropriate assumptions [1], we can use the conditional average treatment effect (CATE) as an estimator for the uplift, where CATE is defined as,

\[
\tau(x) = \mathbb{E}(Y(1) \mid X = x) - \mathbb{E}(Y(0) \mid X = x) = \mathbb{E}(Y \mid T = 1, X = x) - \mathbb{E}(Y \mid T = 0, X = x). \tag{3}
\]

Intuitively, the desired objective can be described as the difference between two conditional means \( \tau(x) = \mu_1(x) - \mu_0(x) \).

III. METHODOLOGY

A. Architecture

The proposed robustness-enhanced uplift modeling framework with adversarial feature desensitization, or RUAD for short, is shown in Fig. 2. Given a sample \( \{x, t, y\} \), where \( x \) may include the categorical features or the numerical features, and the categorical features will be converted into low-dimensional dense vectors through an encoding layer. The feature selection module with joint multi-label training will then train a masker \( m(x) \) with the temperature weights on all the input features to filter the desired key features. We will also adopt a joint optimization involving the true response \( y \) and the transformed response \( y^* \) as a trade-off supervision objective of the process to further ensure that the selected features conform to the desired properties. After receiving the sample \( x_{\text{adv}} \) after feature selection, the adversarial feature desensitization module first uses an adversarial training operation to obtain the adversarial example \( x_{\text{adv}} \) with the largest perturbation on the key feature level and then uses the soft interpolation operation to combine it with the masked sample to obtain a milder adversarial example \( \tilde{x}_{\text{adv}} \). Finally, we jointly train the model...
with the adversarial samples and observed sample sets to free
the model from feature sensitivity, and the final optimization
objective function of our RUAD can be expressed as follows,
\[
\min_{\theta} \mathcal{L}_{RUAD} = \mathcal{L}_o + \mathcal{L}_r + \mathcal{L}_a + \lambda\|\theta\|, \quad (4)
\]
where \(\mathcal{L}_o\), \(\mathcal{L}_r\), and \(\mathcal{L}_a\) denote the prediction loss of the trasformed
and true responses for the feature selection module, and the adversarial loss for the adversarial feature desen-
sitzation module, respectively. \(\lambda\) and \(\|\theta\|\) are the trade-off
parameters and the regularization terms. Next, we describe
each module in detail based on the training process.

B. Base Model

Since our RUAD is model-agnostic, it can be integrated
with most existing uplift models. For the convenience
of description, we use S-Learner [3] as the base model for
an example, but different uplift models will be integrated into
the experiments to verify the compatibility of our RUAD.
In S-Learner, the samples from the treatment group and the
control group will be distinguished by a group index and
trained through a shared base learner. During inference,
the group index of each sample will be modified to obtain another
counterfactual conditional mean function required in Eq.(3).

C. Feature Selection Module with Joint Multi-Label Modeling

Based on our experimental findings in Fig. 1, it can be
found that the perturbation of only some of the key features
will cause significant performance changes in the uplift model,
\(i.e.,\) not all the features have sensitivity problems. Therefore,
to solve the feature sensitivity problem suffered by the uplift
model, we first need to enable the uplift model to identify
the key sensitive features we expect to obtain from all the
input features. This involves a feature selection process and a
reasonable optimization objective that guides the process
to perform correctly. Based on this idea, our RUAD first
formulates a feature selection module with joint multi-label
training. Note that although feature selection has been exten-
sively studied in different research areas, including different
designs for selection strategies or guidance objectives [7], it
is still less addressed in uplift modeling.

1) Feature Selection Process: In this paper, we utilize
a neural network-based masking function, denoted as \(\hat{m}(\cdot)\),
to determine the contribution of each feature in the uplift
modeling. An additional Gumbel-Softmax trick [8] is used
to constrain the model to obtain an approximate \(k\)-hot mask
vector \(m(x)\). The \(k\)-ratio features with the largest contribution
are regarded as desired key features, while the rest are regarded
as irrelevant or redundant ones. The \(m(x)\) is formulated as,
\[
m(x) = \text{Gumbel-Softmax}(\hat{w}(x), kN) \in \mathbb{R}^N, \quad (5)
\]
where \(k = \lfloor kN \rfloor \in \mathbb{Z}_+\) is the number of features expected to
be obtained, \(N\) is the number of input features. Specifically,
In Eq.(5), let \(z = \hat{w}(x) \in \mathbb{R}^N\) be a probability vector, then for
any feature dimension \(j \in \{1, \ldots, N\}\), we have \(z_j \geq 0\) and \(\sum_j z_j = 1\). Based on a pre-defined temperature weight \(\zeta > 0\),
the calculation of each feature dimension in the mask vector
can be expressed as [9],
\[
m_j = \max_{l \in \{1, \ldots, k\}} \frac{\exp\left(\log z_j + \xi_j^l / \zeta\right)}{\sum_i^N \exp\left(\log z_j^i + \xi_j^i / \zeta\right)}, \quad (6)
\]
where \(l \in \{1, \ldots, k\}\) denotes the index of the selected feature,
\(\xi_j^l = -\log(-\log u_j^l)\), and \(u_j^l \sim \text{Uniform}(0, 1)\) denotes
a uniformly distributed sampling. Note that for simplicity, we
follow the setup of previous work [9], \(i.e.,\) \(\zeta = 0.5\). Finally,
we can get the masked samples \(x_m\) by multiplying the original
samples \(x\) with the resulting mask vector \(m(x)\),
\[
x_m = x \odot m(x), \quad (7)
\]
where \(\odot\) denotes the element-wise multiplication.

2) Joint Multi-Label Modeling: The success of the feature
selection process largely depends on a reasonable guiding opti-
imization objective. Most of the existing uplift models adopt
a traditional optimization objective for response modeling,
which directly constrains the model to fit the true response
\(y\) of each sample,
\[
\mathcal{L}_r = \mathcal{L}(y(x), y(l)) \quad (8)
\]
This can ensure the coherent prediction of the model on the
user response. However, we can find that it is not consistent
with the desired objective (i.e., Eq.(3)), and this will make the performance of the uplift model easily uncontrollable.

On the other hand, there is little work focusing on establishing the link between the user responses $y$ and the expected uplift effect $\tau$ [10], among which the transformed response is one of the most representative ways. The specific form of the transformed response is shown in Eq.(9).

$$y^* = \frac{y}{\pi(x)} \cdot t - \frac{y}{1 - \pi(x)} \cdot (1 - t),$$  \hspace{1cm} (9)

where $\pi(x)$ is the propensity score estimation function and is usually modeled by a neural network in practice. The Eq.(9) transforms the observed true response $y$ into $y^*$, such that the expected uplift predictions $\tau$ equals the conditional expectation of the transformed response $y^*$. Since the transformed response is a consistent unbiased estimator of the uplift effect $\tau$, we can fit it with the uplift prediction of the base model to improve the base model’s ability to capture the uplift effect,

$$\hat{y}_x^* = \mu_1(x) - \mu_0(x).$$  \hspace{1cm} (10)

However, using only Eq.(10) as the objective may also cause the predicted response of the model to violate the true response of the user, i.e., damaging the coherent perspective. Therefore, to better guide the training of the above feature selection process to obtain the desired key features that have a greater impact on the performance of the uplift model, we define a joint multi-label trade-off optimization objective,

$$L_o + L_r = \alpha L(\hat{y}_x^*, y^*) + (1 - \alpha) L(m(x), y(t)),$$  \hspace{1cm} (11)

where $\alpha$ is the loss weight, and note that each prediction is based on masked samples $m(x)$ after feature selection.

D. Adversarial Feature Desensitization Module

After obtaining the desired features, the next key step is how to effectively reduce the sensitivity of the uplift model to these sensitive features during its training process. Based on the empirical findings in Fig. 1, we find that the feature sensitivity is reflected in the inadaptability of the uplift model to the perturbations on these key features. Given that existing works [11] show that adversarial training with feature importance can effectively address the limitation of adversarial non-robust features, our RUAD formalizes an adversarial feature desensitization module, including an adversarial training operation and a soft interpolation process.

1) Adversarial Training Operation: In this paper, we follow the virtual adversarial training framework (VAT) [12] to obtain ideal adversarial samples. Specifically, to strengthen the interference of adversarial samples on the model’s uplift effect estimation, we first modify the original adversarial loss to a form based on the transformed response,

$$
\max_{x_{adv}} L(\hat{y}_{adv}^*, y^*), 
$$  \hspace{1cm} (12)

where $\hat{y}_{adv}^*$ is estimated by using $x_{adv}$ as input in Eq.(10). Then, we perform the search process based on the power iteration method proposed by VAT, where new adversarial samples obtained at each iteration are calculated as follows,

$$x_{adv}^{(z+1)} = x_{adv}^{(z)} + \epsilon \cdot \nabla_{x_{adv}} \mathcal{L}(\hat{y}_{adv}^*, y^*) \odot m(x),$$  \hspace{1cm} (13)

where $\epsilon$ is a hyper-parameter to control the step size of the perturbation, and $z$ is the number of iterations. Note that we will use the masked samples after feature selection as the initialization of this search process, i.e., $x_{adv}^{(0)} = x_m$, and apply the same mask $m(x)$ to the perturbations to ensure that adversarial training is only performed on selected key features.

2) Soft Interpolation Process: Since the adversarial training operation and the feature selection module are jointly trained, excessively large perturbations on some features in the early stages of model training may damage the effect of feature selection. To control the magnitude of the adversarial perturbation within a more moderate level, we integrate the obtained adversarial examples (i.e., $x_{adv}^{(0)}$) and the received original samples (i.e., $x_{adv}$ or $x_m$) in a soft interpolation form. Specifically, the final adversarial examples can be obtained as follows,

$$x_{adv} = \gamma \cdot x_{adv}^{(0)} + (1 - \gamma) \cdot x_{adv}^{(Z)},$$  \hspace{1cm} (14)

where $\gamma \sim \text{Uniform}(0, 1)$, and $Z$ is the number of iterations for the power iteration method. After obtaining ideal adversarial examples, we expect the uplift model to adapt to them during training. Therefore, we introduce an adversarial loss to constrain the model not to having large prediction differences between the adversarial examples and the original samples,

$$L_o = \beta L(\hat{y}_{adv}^*, y^*),$$  \hspace{1cm} (15)

where $\beta$ is the adversarial loss weight.

IV. EXPERIMENTS

A. Experiment Setup

1) Datasets: To compare the model performance from an uplift ranking perspective, we use two datasets to show the effectiveness of our training framework: 1) IHDP [13]. The IHDP dataset is utilized as a semi-synthetic dataset to assess predicted uplift. This evaluation involves the synthetic generation of counterfactual outcomes based on the original features, along with the introduction of selection bias. The resulting dataset contained 747 subjects (608 control and 139 treated) with 25 features (6 continuous and 19 binary features) that described both the characteristics of the infants and the characteristics of their mothers. For such kind of short video platforms, clarity is an important user experience indicator. A decrease in clarity may lead to a decrease in users’ playback time. Therefore, through random experiments within a week, we provided high-clarity videos ($T = 1$) to the treatment group, and low-clarity videos ($T = 0$) to the control group. We count the total viewing time of users’
We compare RUAD with T-Learner (T-NN) following the setup of previous work [3], Causal Forest [14], their effectiveness of our RUAD, where the carefully designed two modules can effectively collaboratively discover sensitive key features and perform adversarial feature desensitization to improve the model’s performance.

C. Ablation Study

Next, we conduct the ablation studies of our RUAD and analyze the role played by each module. We sequentially remove the two components of the RUAD, i.e. the feature selection module (FS) and the adversarial feature desensitization module (AFD). We construct three variants of RUAD, which are denoted as RUAD (w/o FS-JMM), RUAD (w/o FS), and RUAD (w/o AFD). Note that RUAD (w/o FS-JMM) represents that we only use the response as the training label of the base uplift model. We present the results in Table II and we can see that removing any part may bring performance degradation.

D. Robustness Evaluation

To verify the effectiveness of our RUAD on different uplift models, except for S-NN, we also combine it with two typical

<table>
<thead>
<tr>
<th>Methods</th>
<th>IHDP Dataset</th>
<th></th>
<th></th>
<th></th>
<th>Production Dataset</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>q (10 bins)</td>
<td>p (15 bins)</td>
<td>q (10 bins)</td>
<td>p (10 bins)</td>
<td>q (10 bins)</td>
<td>p (15 bins)</td>
<td>q (10 bins)</td>
<td>p (10 bins)</td>
<td></td>
</tr>
<tr>
<td>S-NN</td>
<td>0.5455 ± 0.0698</td>
<td>0.4879 ± 0.0784</td>
<td>0.5071 ± 0.0674</td>
<td>0.4576 ± 0.0643</td>
<td>1.2123 ± 0.0104</td>
<td>0.4424 ± 0.0122</td>
<td>1.1766 ± 0.0076</td>
<td>0.3987 ± 0.0156</td>
</tr>
<tr>
<td>Causal Forest</td>
<td>0.7991 ± 0.0002</td>
<td>0.8204 ± 0.0001</td>
<td>0.8185 ± 0.0003</td>
<td>0.7994 ± 0.0002</td>
<td>1.7244 ± 0.0056</td>
<td>0.7666 ± 0.0231</td>
<td>1.8102 ± 0.0064</td>
<td>0.5988 ± 0.0133</td>
</tr>
<tr>
<td>TO-SN</td>
<td>0.8301 ± 0.0922</td>
<td>0.7944 ± 0.0913</td>
<td>0.8233 ± 0.0981</td>
<td>0.8102 ± 0.0897</td>
<td>2.1798 ± 0.0089</td>
<td>0.7666 ± 0.0113</td>
<td>2.3100 ± 0.0075</td>
<td>0.6388 ± 0.0121</td>
</tr>
<tr>
<td>TARNet</td>
<td>0.7233 ± 0.1022</td>
<td>0.7603 ± 0.0605</td>
<td>0.7408 ± 0.0985</td>
<td>0.7790 ± 0.0806</td>
<td>0.7950 ± 0.0051</td>
<td>0.3454 ± 0.1615</td>
<td>0.7790 ± 0.0006</td>
<td>0.3089 ± 0.0145</td>
</tr>
<tr>
<td>CFR</td>
<td>0.7487 ± 0.0893</td>
<td>0.7463 ± 0.0703</td>
<td>0.7463 ± 0.0703</td>
<td>0.7291 ± 0.0699</td>
<td>0.9466 ± 0.0477</td>
<td>0.4666 ± 0.0137</td>
<td>0.8996 ± 0.0099</td>
<td>0.5677 ± 0.0102</td>
</tr>
<tr>
<td>CFRdirect</td>
<td>0.7396 ± 0.0912</td>
<td>0.7542 ± 0.0851</td>
<td>0.7782 ± 0.0925</td>
<td>0.7298 ± 0.0945</td>
<td>0.9608 ± 0.0338</td>
<td>0.6785 ± 0.0228</td>
<td>1.0452 ± 0.0088</td>
<td>0.6887 ± 0.0152</td>
</tr>
<tr>
<td>Dragonnet</td>
<td>0.8324 ± 0.0721</td>
<td>0.8094 ± 0.0842</td>
<td>0.8306 ± 0.0795</td>
<td>0.8577 ± 0.0844</td>
<td>1.6453 ± 0.0102</td>
<td>0.4999 ± 0.0254</td>
<td>1.8308 ± 0.0071</td>
<td>0.6544 ± 0.0164</td>
</tr>
<tr>
<td>CITIE</td>
<td>0.8099 ± 0.1120</td>
<td>0.7996 ± 0.0839</td>
<td>0.8277 ± 0.0742</td>
<td>0.7893 ± 0.1198</td>
<td>0.8467 ± 0.0108</td>
<td>0.5233 ± 0.0146</td>
<td>0.8866 ± 0.0121</td>
<td>0.6017 ± 0.0221</td>
</tr>
</tbody>
</table>

We present the comparison results of IHDP and Production datasets in Table I, and we can observe that our RUAD outperforms our baselines in most cases. Note that we use $\tilde{q}$ as the reference for hyperparameter search, and have a significant advantage on this metric while maintaining a competitive result on other metrics. This demonstrates the effectiveness of our RUAD, where the carefully designed two modules can effectively collaboratively discover sensitive key features and perform adversarial feature desensitization to improve the model’s performance.

E. Compatibility Evaluation (RQ4)

To verify the effectiveness of our RUAD on different uplift models, except for S-NN, we also combine it with two typical

---

Authorized licensed use limited to: SHENZHEN UNIVERSITY. Downloaded on March 08, 2024 at 05:16:39 UTC from IEEE Xplore. Restrictions apply.
models, i.e., T-NN, and Dragonnet, in our experiments. The results of the Production datasets are shown in Fig. 4. We can find that integrating our RUAD on different uplift models can always achieve a performance improvement. This suggests that our RUAD can serve as a general framework to improve the accuracy and robustness of uplift models.

**V. RELATED WORKS**

Uplift modeling has received much attention for online marketing in recent years [1]. Research in this area has focused on various aspects of uplift modeling, including methods for model building, performance evaluation, and real-world applications. For binary outcome, the intuitive approach to model uplift is to build two classification models [3]. This consists of two separate conditional probability models, one for the treated users, and the other for untreated users. This method is simple and flexible, but it can not mitigate the influence of disparity in feature distributions between treatment and control groups. To directly model the uplift, a transformed response approach [10] is proposed, but it heavily relies on the accuracy of the propensity score. For continuous outcome, Causal Forest [14] is a random forest-like algorithm for uplift modeling. It uses the causal tree as its base learner, which is a general framework with theoretical guarantees. With the development of deep learning in causal inference, there are many works proposed that focus on ITE estimation. TARNet [15] is a two-head structure like T-Learner, but the information between two heads is shared by a representation layer. CFRNet leverages the distance metrics (MMD and WASS) based on the structure of TARNet to balance the representation between the two heads. To solve the sample imbalance between the treatment and control groups, Dragonnet [16] designs a tree-head structure, which uses a separated head to learn the propensity score. The propensity score is commonly used in ITE estimation. CITE [17] uses it to distinguish the positive and negative samples, and then builds a contrastive learning structure. Unlike the above methods, our RUAD builds the transformed outcome and conditional means together, which can leverage the deep neural network to obtain better feature representations for uplift modeling.

**VI. CONCLUSION**

In this paper, to address the feature sensitivity problem existing in most uplift modeling methods, we propose a robust-enhanced uplift modeling framework with adversarial feature desensitization (RUAD). RUAD consists of two customized modules: 1) the feature selection module with joint multi-label modeling selects the key sensitive features for the uplift prediction, which can help get a more accurate and robust uplift prediction; and 2) the adversarial feature desensitization module adding perturbations on the key sensitive features can help solve the feature sensitivity problem. We conduct extensive evaluations to validate the effectiveness of RUAD and demonstrate its robustness to the feature sensitivity issue and the compatibility with different uplift models.

**REFERENCES**


