Reproduction Of The Uplift Modeling Combining Expert Networks And Reparameterization Techniques

Abstract

Uplift Modeling is a technique for predicting individual behavior changes following a specific intervention, and it is widely applied in online marketing to help businesses optimize resource allocation and improve return on investment (ROI). Currently, the primary methods include Meta-learning-based approaches, tree/-forest models, and neural networks; however, these methods primarily focus on binary treatment scenarios. In addressing multi-valued treatment scenarios, this paper, inspired by M^3TN and EFIN, proposes a novel model that combines multi-gate expert networks with an explicit modeling module for uplift effects, and its superiority is validated on the Criteo dataset. The results demonstrate that the proposed model achieves high efficiency and accuracy in both multi-valued and binary treatment contexts, providing an effective solution to causal inference problems in online marketing.

Keywords: Uplift Modeling, Individual Treatment Effect, Binary Treatment, Expert Networks.

1 Introduction

Uplift Modeling [2] is a technique focused on predicting individual behavior changes following specific interventions (such as advertising campaigns, discount offers, etc.), and it holds significant application value in the field of online marketing [5]. This technique helps businesses identify target users who are most likely to respond positively to the intervention, thereby enabling precise resource allocation and improving return on investment (ROI). However, in practice, it is impossible to observe the behavior of the same user under different treatment conditions simultaneously. This missing potential outcome in causal inference leads researchers to typically estimate the Individual Treatment Effect (ITE) [12] by indirectly comparing the observed data of the treatment and control groups. Although various Uplift Modeling methods have been proposed, including Meta-learning-based approaches, tree/forest models [12], and neural networks, most of these methods focus on binary treatment scenarios (e.g., presence or absence of an intervention). Research and applications concerning more complex multi-valued treatment scenarios (e.g., interventions of varying intensities) remain underexplored.

In practical applications, traditional binary treatment methods face significant challenges when directly extended to multi-valued treatment scenarios. These challenges typically arise from a decline in modeling efficiency and prediction imbalance between treatment and control groups. Such imbalance may lead to progressively increasing cumulative errors as the treatment value increases, thereby affecting the overall performance of the model. Moreover, the complexity of multi-valued treatment scenarios demands higher modeling accuracy

and efficiency, which existing methods often fail to adequately address. To tackle these issues, recent research has proposed several improved models for multi-valued treatment scenarios, such as the Multi-valued Treatment Network (M^3TN) [11], which demonstrates excellent performance in handling such scenarios. However, the modeling efficiency and predictive performance of these methods on binary datasets have not been systematically evaluated, providing a clear direction for further research.

In this paper, inspired by the multi-valued treatment network based on expert mixture (M^3TN), and drawing on advanced concepts from the Explicit Feature Interaction Perception Network (EFIN) [8] and the Explicit Uplift Effect Network (EUEN) [6], we propose a novel model, namely the Multi-Gate Expert Mixture Network with Explicit Uplift Effect Modeling. By combining the strengths of various expert networks, this model not only improves modeling efficiency but also significantly enhances prediction accuracy. To validate the effectiveness of the proposed model, we conducted experiments on the Criteo dataset and compared it with several existing modeling methods. The experimental results demonstrate that the proposed model not only outperforms existing methods on training data but also exhibits high efficiency and accuracy in practical applications.

2 Related works

Consider an observed dataset $D = \{(x^i, t^i, y^i)\}_{i=1}^n$ consisting of n samples, where $x_i \in R^d$ is a d-dimensional feature vector, $t_i \in \{0, 1, ..., K\}$ is the treatment variable (e.g., different levels of discount), and $y_i \in Y$ the response variable (which can be binary or continuous). In the Neyman-Rubin potential outcomes framework, the potential outcomes for individual i under treatment $t_i = k$ or no treatment $t_i = 0$ are denoted as $y_i(k)$ and $y_i(0)$, respectively. The Individual Treatment Effect (ITE) [12] for multi-valued treatments is defined as:

$$\tau_i^k = y_i(k) - y_i(0),$$

which represents the incremental effect of treatment k on individual i. However, since each individual can only observe one treatment outcome $y_i(k)$ and $y_i(0)$, the true value of τ_i^k cannot be directly observed. Under appropriate assumptions, the Conditional Average Treatment Effect (CATE) [4] can be used as an unbiased estimate of τ_i^k , which is defined as:

$$\tau^k(x) = E[y(k) - y(0) \mid x] = E[y(k) \mid t = k, x] - E[y(0) \mid t = 0, x],$$

where $\mu_k(x) = E[y(k) \mid t = k, x]$ and $\mu_0(x) = E[y(0) \mid t = 0, x]$ represent the conditional means of the treatment group and control group, respectively, given the features x. The ultimate goal is to estimate $\hat{\tau}^k(x) = \mu_k(x) - \mu_0(x)$ to predict the uplift effect for each individual, and based on these uplift values, rank the users to guide treatment allocation decisions.

3 Method

3.1 Overview

This method combines the Mixture of Experts layer(MoE) [3] with a reparameterization module to construct an efficient neural network architecture aimed at enhancing feature learning and inference efficiency. In the feature representation phase, the MoE layer collaboratively leverages multiple expert networks and gating networks to achieve fine-grained modeling of the feature subspace. The sparse activation mechanism significantly improves computational efficiency while maintaining high capacity. Building upon this, the reparameterization module increases the model's expressive power and generalization capability during training by enhancing network complexity, while simplifying the network structure during inference through parameter reorganization to reduce computational overhead. The synergy of these two components achieves an optimized balance between performance and efficiency in complex tasks, demonstrating superior application potential.

3.2 Feature Representation Module

The Mixture of Experts (MoE) layer [3] is composed of multiple expert models and a Gating Network [9]. These expert models are parallel subnetworks, each specialized in processing different feature subspaces, thereby leveraging domain-specific expertise. The gating network dynamically assigns weights based on the input features, intelligently selecting the most relevant expert models for activation. This mechanism, through sparse activation, ensures that only a few expert models are activated during data processing, significantly reducing computational resource consumption. Despite this, the MoE layer retains its high-capacity nature, as it combines the activated expert models' outputs with weighted aggregation through the gating network, ultimately producing the MoE layer's comprehensive output.

Then, the feature representation ϕ_k of each prediction head can be formulated as:

$$\phi_k(x) = \sum_{n=1}^{N} g_k(x) f_n(x), k \in \{0, 1, ..., K\},$$

where f_n represents the expert layer,the gating networks g_k are simply linear transformations of the input with a softmax layer:

$$g_k(x) = softmax(W_{gk}x)$$

3.3 Reparameterization Module

The Reparameterization Module is a technique designed to optimize the training and inference performance of neural networks through structural transformation. The core idea behind this technique is that, during the training phase, a more complex structure is employed to enhance the model's expressive capacity, allowing it to capture subtle features and patterns within the data. This approach enables the model to achieve higher accuracy and generalization during the learning process. In the inference phase, when the model is required to make predictions or classifications on new data, the Reparameterization Module simplifies the network structure through a parameter reorganization strategy. This strategy aims to reduce the computational overhead of

the model in practical applications, thereby improving operational efficiency and ensuring that the model can perform inference tasks quickly and accurately in resource-constrained environments, such as mobile devices or embedded systems. Through structural optimization in both training and inference stages, the Reparameterization Module not only enhances the performance of neural networks but also optimizes their practicality in real-world applications.

Then, the reparameterization module can be formulated as:

$$\mu_0(x) = h_0(\phi_0),$$

$$\hat{\tau}^k(x) = h_k(\phi_k), k \in \{1, 2, ..., K\}.$$

where $\mu_0(x)$ represents the control response prediction, and $\hat{\tau}^k(x)$denotes the uplift prediction. Subsequently, $\mu_k(x)$ can be estimated using the additive function $\mu_k(x) = \mu_0(x) + \hat{\tau}^k(x)$.

3.4 Loss

The loss function $J(\theta)$ is for learning the conversion of treatment group, and $L(\theta)$ is the loss function for learning conversion of control group.Specifically:

$$J(\theta) = \sum_i t_i[y_i - (\tau(x_i) + \mu_c(x_i))]^2,$$

$$L(\theta) = \sum_i (1 - t_i)[y_i - \mu_c(x_i)]^2.$$

The optimization process jointly train two sub-networks:

$$\hat{\theta} = \arg\min_{\theta} \ J(\theta) + L(\theta)$$

4 Implementation details

4.1 Experimental environment setup

Datasets. Experiments were conducted on the public Criteo dataset, which contains a subset of traffic data collected by Criteo over a 7-day period, aimed at binary classification of advertisement click-through rates. The dataset includes 40 million training samples and 6 million test samples, with positive (click) and negative (non-click) examples subsampled at different rates to optimize the dataset size. Each instance represents an ad impression event, with features divided into 13 continuous numerical features (primarily based on counts) and 26 categorical features, which have been anonymized using 32-bit hashing. The test set corresponds to events occurring the day after the training period, and a sample submission file is provided for evaluation.

Baselines. To evaluate the effectiveness of M^3TN ,we compare the performance of M^3TN with S-Learner [7], T-Learner [7], TARNet [9], DragonNet [10].

Evaluation Metrics. This study follows the framework of previous research and adopts several commonly used evaluation metrics in uplift modeling, including the uplift of the top k observations based on overall sample uplift(u_at_k), the area under the normalized Qini curve based on predicted scores(qini_coef) [1], the area under the normalized uplift curve based on predicted scores(uplift_auc) [1] , and Weighted Average Uplift(wau).

Hardware. In this experiment, a high-performance computing cluster equipped with 8 V100 GPUs was used for model training.

System Environment. All baselines and M^3TN were implemented using PyTorch 1.13.1, with the Optuna hyperparameter optimization package employed to search for the best parameters for both the baseline models and M^3TN.

Hyperparameter. The model studied in this paper treats the number of convolutional layers, batchsize, learning rate, and the number of experts as hyperparameters.

4.2 Model enhancement

In the current wave of technological innovations, a significant shift has occurred in the realm of reparameterization modules. By adopting binarization techniques, it is now possible to accomplish tasks that previously required more resources and complex structures, using only a single control network and a corresponding uplift network.

Building upon the M^3TN framework, the reparameterization module has been modified, drawing inspiration from the Explicit Feature Interaction Network model(EFIN) [8] and the Explicit Uplift Effect Network (EUEN) [6]. The multi-head output in the expert network layer has been transformed into a dual-head output, enabling its application in binary networks. The original model is depicted in Figure 1, while the improved model is shown in Figure 2.

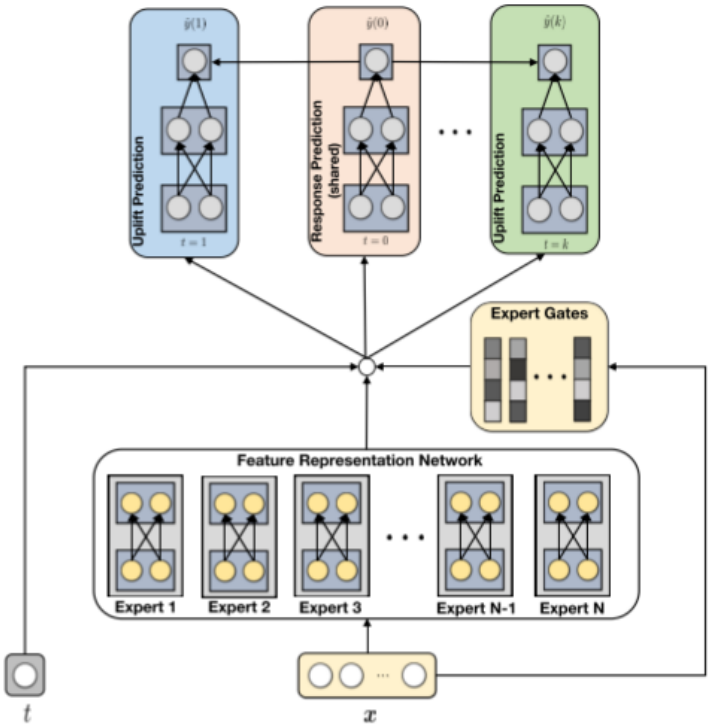

Figure 1. The architecture of the Mixture-of-Experts based Multi- valued Treatment Network (M^3TN)

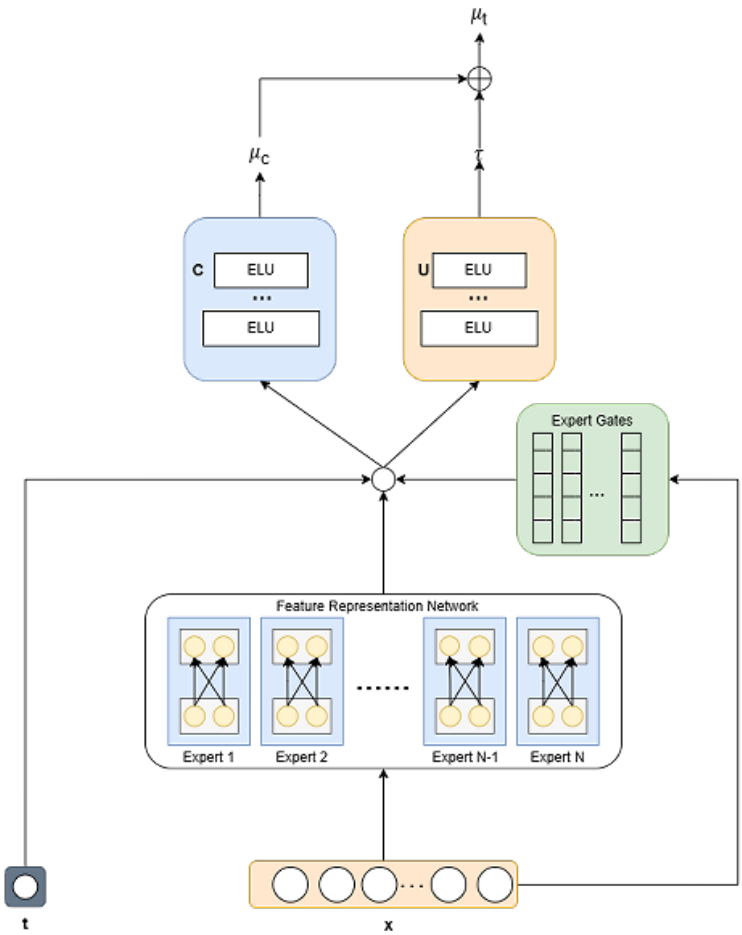

Figure 2. Combining Expert Networks with Reparameterization Techniques for Predicting Uplift in binary scenarios

5 Results and analysis

5.1 Overall Performance

Table 1 presents a detailed comparative analysis of the Criteo dataset, from which several key observations and conclusions can be drawn. First, the T-learner outperforms the S-learner on the Criteo dataset, demonstrating its significant advantage in causal inference tasks. Despite some baseline models utilizing more complex network architectures, the T-learner remains highly competitive across multiple performance metrics, further validating the robustness and reliability of its approach.

Second, the introduction of more complex neural network structures, such as TARNet and DragonNet, facilitates further improvements in model performance. Among these, DragonNet stands out due to its innovative incorporation of a regularization mechanism, particularly excelling in the Weighted Average Uplift (WAU) metric. This suggests that DragonNet is more effective in estimating causal effects, thereby achieving a leading position in key performance indicators.

Finally, compared to other baseline methods, the proposed M^3TN model significantly outperforms all baseline models in most experimental scenarios, showcasing its exceptional performance on the Criteo dataset. The M^3TN model demonstrates substantial advantages in capturing complex data characteristics and enhanc-

ing generalization capability, providing strong evidence of its potential and application value as an advanced method for causal effect estimation.

In summary, these findings indicate that the proposed M^3TN model has broad applicability in causal inference tasks, offering a novel and effective solution for improving the accuracy of causal effect estimation.

5.2 Ablation Studies

Building on the insights from the original study, this research systematically conducts an ablation analysis of the M^3TN model, aiming to explore the specific contributions of its components to the overall model performance. To this end, two key components of the M^3TN , namely the Mixture of Experts (MoE) in the feature representation module and the Reparameterization Module (RM), were sequentially removed, resulting in two variant models, denoted as M^3TN (w/o MoE) and M^3TN (w/o RM). A comparative analysis of the experimental performance of these two variants further reveals the functionality and importance of each module.

The results, as shown in Table 2, are somewhat surprising: after removing the MoE module, the overall performance of the model showed a certain degree of improvement. This finding contrasts sharply with the conclusions presented in the original study, which stated that the removal of either the MoE module or the RM module would result in a performance decline. However, in this study, the removal of the MoE module seemed to have a particularly unique effect on performance. This phenomenon suggests that the MoE module may play a more positive role in multi-valued multi-task learning scenarios, whereas its contribution might be more limited in binary tasks and, under certain conditions, could even constrain performance.

Further analysis indicates that the core mechanism of the MoE module lies in allocating adaptive expert networks to different tasks, thereby enhancing the model's expressive capability in multi-task learning. However, in binary task scenarios, where task complexity and interaction are relatively low, the MoE module may not have fully exploited its potential advantages. This finding provides a new perspective on understanding the performance of the M^3TN model under different task settings, while also highlighting the need for further research into the coupling between task characteristics and module design.

Overall, this study, through the ablation analysis of the M^3TN model, reveals the varying applicability of the MoE module across different task scenarios. This not only provides a basis for a deeper understanding of the internal mechanisms of the M^3TN model but also offers important insights for the future design of more task-specific causal inference models.

Table 1. Overall comparison between our M^3TN and the baselines on Criteo dataset, where the best and second best results are marked in bold and underlined, respectively.

Method	ut_at_k	qini	uplift_auc	wau
s-learner	<u>0.0328</u>	<u>0.0857</u>	<u>0.0332</u>	0.0092
t-learner	0.0393	0.1015	0.0402	0.0093
tarnet	0.0402	0.0997	0.0394	<u>0.0089</u>
dragonnet	0.0379	0.1036	0.0411	**0.0096**
m^3tn	**0.0421**	**0.1051**	**0.0418**	0.00904

Table 2. Results of the ablation studies on the Criteo dataset, where the best and second best results are marked in bold and underlined, respectively.

Method	ut_at_k	qini	uplift_auc	wau
m^3tn (w/o MoE)	**0.0454**	**0.11199**	**0.0447**	**0.0106**
m^3tn (w/o RM)	<u>0.0399</u>	<u>0.1032</u>	<u>0.0409</u>	<u>0.0084</u>
m^3tn	0.0421	0.1051	0.0418	0.00904

6 Conclusion and future work

During the reproduction of the M^3TN model, the researchers enhanced their programming skills and deepened their understanding of the MoE (Mixture of Experts) module, reparameterization module, as well as the EFIN and EUEN models. Experimental validation demonstrated that the M^3TN model exhibited significant performance advantages in multi-valued treatment tasks. Its core components include the feature representation module based on the multi-gate mixture of experts (MoE) and the reparameterization module, which consists of a control network and a gain network. However, ablation experiments revealed that the MoE module contributed only marginally in binary tasks, indicating the model's limited applicability in low-complexity scenarios.

To address this limitation, future research could optimize the learning framework to better suit binary tasks, design lightweight feature modules to improve efficiency, and extend the model to multi-task network scenarios to enhance generalization capability. Additionally, improving the design of the MoE module to accommodate the requirements of both binary and multi-valued tasks would further enhance model performance, providing new theoretical and practical support for multi-valued treatment modeling.

References

[1] Lu Cheng, Ruocheng Guo, Raha Moraffah, Paras Sheth, K Selçuk Candan, and Huan Liu. Evaluation methods and measures for causal learning algorithms. *IEEE Transactions on Artificial Intelligence*, 3(6):924–943, 2022.

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

[3] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[4] Masahiro Kato and Masaaki Imaizumi. Cate lasso: conditional average treatment effect estimation with high-dimensional linear regression. *arXiv preprint arXiv:2310.16819*, 2023.

[5] Shogo Kawanaka and Daisuke Moriwaki. Uplift modeling for location-based online advertising. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Recommendations, Geosocial Networks and Geoadvertising*, LocalRec '19, New York, NY, USA, 2019. Association for Computing Machinery.

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

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

[8] Dugang Liu, Xing Tang, Han Gao, Fuyuan Lyu, and Xiuqiang He. Explicit feature interaction-aware uplift network for online marketing. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4507–4515, 2023.

[9] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

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

[11] Zexu Sun and Xu Chen. M 3 tn: Multi-gate mixture-of-experts based multi-valued treatment network for uplift modeling. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5065–5069. IEEE, 2024.

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