Uplift Modeling based on Graph Neural Network Combined with Causal Knowledge

1st Haowen Wang
Zhejiang Lab
Shanghai, China
wanghw@zju.edu.cn

2nd Xinyan Ye
Imperial College London
London, UK
xy2119@ic.ac.uk

3rd Yikang Wang
University College London
London, UK
yikang.wang.21@ucl.ac.uk

4th Yangze Zhou
Zhejiang University
Hangzhou, China
yangze.zhou@zju.edu.cn

5th Zhiyi Zhang
Peking University
Beijing, China
emma0302@pku.edu.cn

6th Longhan Zhang
Zhejiang Lab
Hangzhou, China
Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
longhanz@zhejianglab.com

**Corresponding author 7th Jing Jiang
Zhejiang Lab
Hangzhou, China
jiangj@zhejianglab.com

††Corresponding author 8th Yiteng Zhai
Zhejiang Lab
Hangzhou, China
Nanyang Technological University
Singapore
ito@zhejianglab.com

***Abstract*—Uplift modeling is a fundamental component of marketing effect modeling, which is commonly employed to evaluate the effects of treatments on outcomes. Through uplift modeling, we can identify the treatment with the greatest benefit. On the other side, we can identify clients who are likely to make favorable decisions in response to a certain treatment. In the past, uplift modeling approaches relied heavily on the difference-in-difference (DID) architecture, paired with a machine learning model as the estimation learner, while neglecting the link and confidential information between features. We proposed a framework based on graph neural networks that combine causal knowledge with an estimate of uplift value. Firstly, we presented a causal representation technique based on CATE (conditional average treatment effect) estimation and adjacency matrix structure learning. Secondly, we suggested a more scalable uplift modeling framework based on graph convolution networks for combining causal knowledge. Our findings demonstrate that this method works effectively for predicting uplift values, with small errors in typical simulated data, and its effectiveness has been verified in actual industry marketing data.**

***Index Terms*—Uplift Modeling, Graph Neural Network, Causal Inference**

I. Introduction

Uplift modeling [1] has traditionally relied on randomized experiments, such as randomized controlled trials (RCTs) [2], in which customers are randomly allocated to either receive or not receive the intervention. In such instances, obtaining an accurate and interpretable estimate from observational data becomes critical. However, carrying out such an experiment in a business context frequently results in several challenges, including high costs in terms of time and money, uneven intervention distribution, and selection bias in the specific population.

Response modeling or outcome prediction uses supervised learning models to model the relation between features and target variables to predict response variation. Although response modeling is typically preferable to random targets, distinguishing between treatment-induced be- behavioral changes is often challenging. The population that should be targeted is the one most likely to respond positively to the intervention. As a result, a thorough knowledge of the behavioral changes that occur after the intervention is essential. Uplift modeling simulates the causal effect between the intervention and the outcomes based on response modeling. Causal inference frameworks and machine learning models are corporated to provide accurate forecasts and optimized performance on intuitive metrics.

The counterfactual nature of intervention data is central to causal inference in Rubin's Potential Outcome Framework (POF) [3]. This characteristic pertains to a person's inability to both receive and refuse intervention. This means that the effects of many therapies cannot be seen in the same person. Two frameworks that have been extensively examined for causal impact estimations based on this counterfactual characteristic are the meta-learner framework [4] and the customised machine learning model-based framework [5]. The ultimate goal is to increase the accuracy of causal impact estimation through the use of feature engineering and validation approaches such

This work was supported by the Key R&D Program of Zhejiang (2024C01036).

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00267

as PS matching [6], weighting [7], feature representation [8], and so on.

In the past, researchers in uplift modeling were largely concerned with how to employ unbiased data and models in the estimation framework. We increased the amount of data information by defining causal knowledge and implementing structured representation, then used a graph convolution neural network [9] to efficiently and directionally integrate feature neighborhood information, achieving excellent performance in uplift modeling and prediction tasks. The following is a description of our paper's contribution to methodological and empirical evaluation perspectives:

• First, we propose to use conditional average treatment effect (CATE) as the attribute representing the causal information of the feature and as part of uplift modeling and propose a causal network model framework to effectively calculate it based on knowledge distilling and double machine learning.

• Second, we propose to learn the causal diagram structure of the data before uplift modeling and reconstruct the data according to the learned adjacency matrix.

• Third, we propose an uplift modeling estimator based on graph convolution neural networks, which can integrate and characterize neighborhood feature attributes according to the cause and effect diagram structure and improve the performance of downstream tasks.

II. Related Work

The estimation of the uplift value in uplift modeling is often based on the Potential Outcome Framework(POF) [3]. The individual treatment effect(ITE) can be expressed as:

$$ITE : \tau(i) = Y_i(1) - Y_i(0) \tag{1}$$

Shere $Y_i(1)$ and $Y_i(0)$ represents the result of the outcome variable under the treatment condition and control condition, respectively, for individual i, $\tau(i)$ is the ITE value.

Considering that the individual effect of treatment will vary from individual and the high cost of marketing experiments in the industry, the conditional average treatment effect(CATE) [10] is proposed as the effect of treatment on subgroups evaluated by the conditional average treatment effect (CATE), which is calculated by:

$$CATE : \tau_i = E\left[Y_i(1) \mid X_i\right] - E\left[Y_i(0) \mid X_i\right] \tag{2}$$

where X_i is the feature vector for individual i.

For the estimation of CATE and ITE, the most direct method is to make an unbiased adjustment to the regression model. Series of meta learners represented by s-learner are designed based on the concept, that is, train one or more models with y as the output training target, input T and X, and get the change of Y by changing the value of T to estimate ITE and CATE.

$$\tau(x) = E[Y_i(1) - Y_i(0)|X] = E[\tau_i|X] \tag{3}$$

Another series of methods for uplift modeling prediction is the probability score matching (PSM) [6] method based on randomized controlled trials (RCT) [2]. By calculating the probability score $P(t \mid x)$, each sample is given a different treatment object according to its similarity, so for sample i, we find sample j:

$$\operatorname{argmin}_j \operatorname{dist}(i, j) = \left|P\left(t \mid x_i\right) - P\left(t \mid x_j\right)\right| \tag{4}$$

Then CATE could be calculated by:

$$\hat{\tau} = \frac{1}{n}\left[\sum_{i:t_i=1}\left(y_i - y_j\right) + \sum_{i:t_i=0}\left(y_j - y_i\right)\right] \tag{5}$$

In addition, the industry's research on lift modeling also includes methods based on the Covariate Balancing Method and Modeling Unobserved Confounder. Typical methods of the first category include Inverse Probability of Treatment Weighting (IPTW) [11], Entropy Balancing (EB) [12], and Approximate Residual Balancing (ARB) [13], in which core is how to re-assign weights to samples. The core of the second type of method is to model the confounder. One way is to model the instrumental variable, which is represented by the two-stage least square (2SLS) method [14]. The first stage is to fit the impact of the change of I on T, and the second stage is to fit the impact of the change of T on y caused by the change of I. The other way is to use deep learning to represent the confounder, such as SITE [15], Dragonnet [16], and CEVAE [17].

The past research mainly focused on adjusting and optimizing the uplift value estimation model in a structured or unstructured way. Estimation methods based on the foundation model have shown us the importance of embedding causal structure knowledge into the estimation process. This paper will try to conduct data mining on features. On the one hand, it expands the amount of information by defining and applying causal information; on the other hand, it reconstructs structured origin data through causal diagram structural information and uses GCN to learn neighborhood information from unstructured reconstructed data to improve the performance of uplift modeling using the framework of meta learner [4].

The structure of the paper is as follows. Section II reviews the critical concepts of uplift modeling and frameworks of the learning approach. In Section III, we introduce the methodology of our causal knowledge framework. Section IV evaluates these methods with both synthetic and real-world data. Finally, Section V summarizes the findings and recommends future research for uplift modeling applications.

III. Methodology

This section will introduce the calculation method and architecture of graph neural networks embedded with causal knowledge. We propose an interpretable causal graph network representation learning framework with features as nodes. It can expand the representation of features by node embedding, mapping the originally scalar features into a high-dimension space, and then integrating the causal information and structural information into the graph features through

graph convolution to achieve a more accurate estimation of uplift value.

A. Causal Knowledge Representation

We propose a framework for computing causal knowledge representation. We transfer knowledge through the concept of the soft target in knowledge distillation as the estimation target of the causal estimator. We estimate each feature's causal average treatment effect(CATE) and take it as the weight of the feature based on the causal effect. This work has been proven to obtain more information.

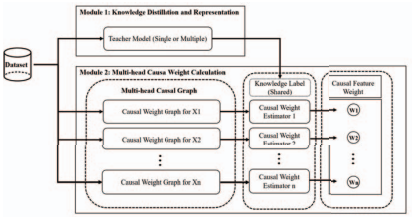

Fig. 1. Causal Weighting Calculation Framework

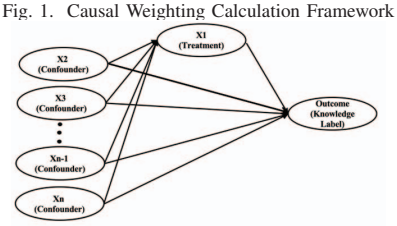

Fig. 2. Causal Graph for Feature X1

Figure 1 shows the architecture of the causal average treatment effect(CATE) calculation. Firstly, in the module of knowledge distillation and representation, We will build a knowledge distillation task for label Y(0/1), using the teacher model(XGBoost [18], etc. as base regressor) to get the probability $\hat{Y}$ as the soft label to replace Y as the target label. Secondly, in the multi-head causal weight calculation module, we establish a causal graph for each feature as shown in Figure 2, use the soft label $\hat{Y}$ got in Module 1 as a knowledge label, and estimate CATE in the framework of double machine learning (DML) [19]. Since the CATE estimation of each feature is independent, we designed a multi-head mechanism to make the calculation more efficient. Double machine learning is a classic estimator to estimate (heterogeneous) treatment effects when treatment is classified and all potential confounders/controls. DML makes the following structural equation assumptions for the data generation process:

$$Y = \theta(X) \cdot T + g(X, W) + \epsilon \quad \mathbb{E}[\epsilon \mid X, W] = 0 \tag{6}$$

$$T = f(X, W) + \eta \quad \mathbb{E}[\eta \mid X, W] = 0 \tag{7}$$

$$\mathbb{E}[\eta \cdot \epsilon \mid X, W] = 0 \tag{8}$$

After modeling Y and T, respectively, the estimated CATE value $\theta(X)$ satisfies the equation:

$$\tilde{Y} = \theta(X) \cdot \tilde{T} + \epsilon \tag{9}$$

Here $\tilde{Y}$ is the residual of Y, $\tilde{T}$ is the residual of T.

Considering $\mathbb{E}[\epsilon \cdot \eta \mid X] = 0$, the problem of estimating $\theta(X)$ can be transformed into the following regression problem.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \mathbb{E}_n \left[(\tilde{Y} - \theta(X) \cdot \tilde{T})^2 \right] \tag{10}$$

B. Causal Graph Structure Learning

The graph network structure contains the dataset's prior information. The connection relationship indicates the direction and distance of information transmission and determines the direction and degree of information sharing and transmission of nodes in the subsequent graph network characterization operation.

Here, we use the classical Bayesian network structure as the structure of the causal feature representation graph. Scoring search is a standard method to solve the problem of Bayesian network structure to evaluate the degree of fit between the Bayesian network and training data and then find the optimal Bayesian network based on the scoring function. The goal is now to solve the following task:

$$\underset{G \in G}{\operatorname{argmax}} \operatorname{score}(G, \mathcal{D}). \tag{11}$$

The scoring function introduces the inductive preference of what kind of Bayesian network you want to obtain. Here we use the Bayesian Information Criterion(BIC) [20] as the score function, which approximates the Bayes Dirichlet equivalent uniform(BDeu), sharing the critical property of decomposability.

$$\operatorname{score}(G, \mathcal{D}) = \sum_{X_i} \operatorname{score}\left(X_i, \Pi_i, \mathcal{D}\right) \tag{12}$$

$$\operatorname{Score}_{bic}(g : D) = l((\hat{\theta}, g) : D) - \frac{\log M}{2} \operatorname{Dim}[g] \tag{13}$$

D is the given data, M is the number of training samples, g is the given structure, Dim[g] is the number of independent parameters of model g, $\hat{\theta}$ is the maximum likelihood estimate of the parameter given the structure g and the data D.

After determining the scoring function, here we use the hill-climbing [21] algorithm as the optimization algorithm for the structural learning problem. The Hill-climbing algorithm is a classical algorithm for local search based on a greedy algorithm, starting with a candidate solution and continuing to search in its neighborhood until there is no better solution. The steps of the local search algorithm are described as follows: Firstly, initialize a feasible solution X. Secondly, select a moved solution s (x) in the neighborhood of the current solution so that $f(s(x)) < f(x), s(x) \in S(x)$. If there is no such solution, X is the optimal solution, and the algorithm stops. Thirdly, make x = s (x) and repeat the second step.

C. GNN based uplift modeling

After Causal Knowledge Representation and Causal Graph Structure Learning, we obtained more information about the dataset and a specific relationship between features. Considering the excellent representation ability of graph neural networks, we propose a graph neural network representation framework based on causal graph representation, which can integrate this information more efficiently.

GCN [9] is a multi-layer neural network that can operate directly on the graph and induce nodes to obtain information on neighborhood vectors based on the neighborhood attributes of nodes. Consider a graph $G = (V, e)$, where $(\|V\| = n)$ and E are the sets of nodes and edges, respectively. It is assumed that each node is connected to itself, that is, (v, v) $\in E$ for any v. Let x $\in \mathbb{R}^{n\times m}$ be a matrix containing all N nodes and their vector features, where m is the dimension of the vector, and each row $x_v \in \mathbb{R}^m$ are the vectors of V. We introduce the adjacency matrix A and its degree matrix D of G, where $d_{ii} = \sum_j A_{ij}$. Due to the characteristics of the self-circulation hypothesis, the diagonal element of a is set to 1. In general, GCN can only capture information about its neighbors through one layer of convolution. We can integrate information about a wider range of neighbors by stacking multiple GCN layers:

$$H^{(l+1)} = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{14}$$

Here $H^{(l+1)}$ and $H^{(l)}$ are the output and input matrices. $\hat{A} = A + I$, where A is the adjacency matrix, and I is the identity matrix. $\hat{D}$ is the degree matrix of $\hat{A}$, $\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix, and $W^{(l)} \in \mathbb{R}^{m\times k}$ is a weight matrix. σ is an activation function, e.g., a $\mathrm{LeakyReLU}$.

Here we use GCN to extract and integrate features based on the causal neighborhood structure we learned in the previous step. We take advantage of the feature that GCN can efficiently fuse features according to the neighborhood structure to get graph embedding of each sample and then perform prediction tasks based on it. Figure 3 shows that we have expanded the information on each feature. In addition to the value of each feature itself, we have also expanded the information of structural features and causal weights.

For the estimation of uplift value, we refer to the design method of S-learner in meta learner and use our GNN-based model as the base learner.

$$\mu_0(x) = \mathbb{E}[Y(T = 0) \mid X = x] \tag{15}$$

$$\mu_1(x) = \mathbb{E}[Y(T = 1) \mid X = x] \tag{16}$$

After μ_0 and μ_1 are calculated, respectively, the uplift value for each sample can be calculated:

$$\tilde{D}_i^1 := Y_i^1 - \hat{\mu}_0\left(X_i^1\right) \tag{17}$$

$$\tilde{D}_i^0 := \hat{\mu}_1\left(X_i^0\right) - Y_i^0 \tag{18}$$

Here $\tilde{D}_i^1$ and $\tilde{D}_i^0$ are the uplift values for samples in the intervention group and control group.

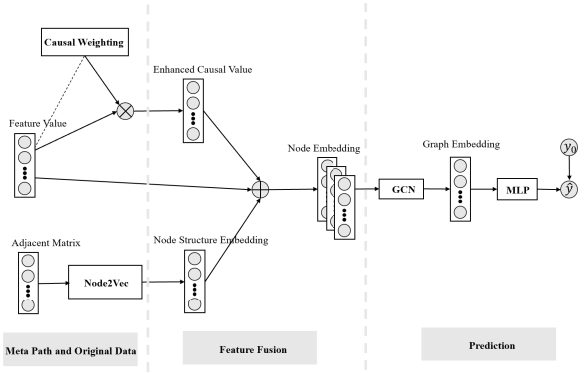

Fig. 3. GNN-based uplift modeling architecture.

IV. Experiments

A. Dataset

1) Synthetic dataset: We used a method to simulate the generation of a dataset containing individual treatment effects, which is available in causalml. In [5] research, it is used as a method to provide simulated data, which is available in Causalml. This synthetic method in the study provides the test groundings for estimating individual treatment effects and facilitating validation. The following is the generating mechanism: for different choices of X-distribution P_d, there is dimension d, noise level σ, propensity function $e^*(\cdot)$, baseline primary effect $b^*(\cdot)$, and treatment effect function $\tau^*(\cdot)$. The distributions and relations are mathematically expressed in terms:

$$X_i \sim P_d \tag{19}$$

$$\varepsilon_i \mid X_i \sim N(0, 1) \tag{20}$$

$$W_i \mid X_i \sim \text{ Bernoulli } \left(e^*\left(X_i\right)\right) \tag{21}$$

$$Y_i = b^*\left(X_i\right) + \left(W_i - 0.5\right)\tau^*\left(X_i\right) + \sigma\varepsilon_i \tag{22}$$

The generative mechanism is characterized by nuisance components and a straightforward treatment effect function. The initial distribution is established with $X_{i1} \sim \mathrm{Unif}(0, 1)^d$, succeeded by the computation of the propensity score $e^(X_i) = \mathrm{trim}0.1\sin(\pi X_{i1}X_{i2})$ and the treatment effect $\tau^(X_i) = \frac{X_{i1}+X_{i2}}{2}$. The treatment variable (W) is subsequently generated as a binary outcome. Finally, interval trimming of the distribution is enacted using the function $\mathrm{trim}(x) = \max\eta, \min(x, 1-\eta)$, where η represents the trimming threshold.

This simulation approach is conceived as a scaled adaptation of the Friedman function [22], wherein a baseline main effect is determined by $b^*(X_i) = \sin(\pi X_{i1}X_{i2}) + 2(X_{i3} - 0.5)^2 + X_{i4} + 0.5X_{i5}$.

2) Real-world dataset: We use the criteo uplift dataset [1] as the evaluation of the real-world dataset, which is constructed by collecting data from the incremental test. It randomly divides the people into two categories, whether it is advertised or not. The criteo uplift dataset has 25 million rows, each representing a user with 11 characteristics, a treatment indicator, and two tags (click and conversion). Here we use conversion as the tag we focus on in uplift estimation. Figure 4 shows the results after learning the Bayesian network structure of the dataset.

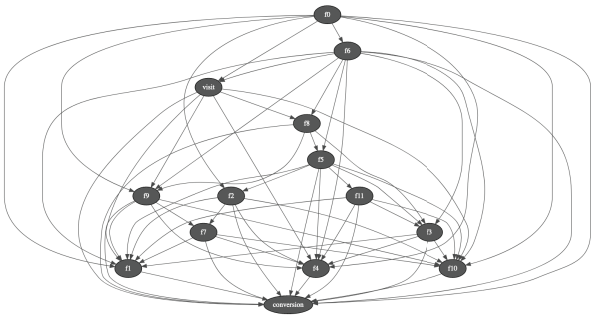

Fig. 4. Correlation network between confounders, treatment, and outcome from the real-world dataset, CRITEO

B. Result

1) synthetic dataset: In this case, the actual causal effect of features can be calculated easily because the datasets are produced with a certain mechanism The absolute loss (*Abs*) is adopted to measure the deviation between the actual causal effect and the estimated causal effect. As for the prediction accuracy, the mean squared error (*MSE*) is adopted. The proposed method has been compared to traditional models like linear regression (LR), SVR, and XGBoost.

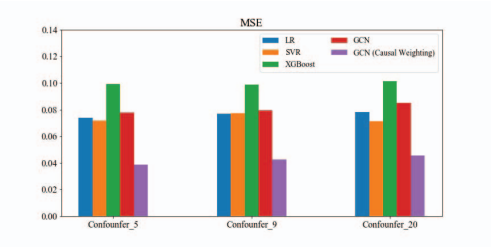

Fig. 5. Mean squared error for y-prediction accuracy of base methods and ours with the numbers of confounders are 5, 9, and 20, respectively.

Figure 5 shows that the origin GNN-based model performs similarly to traditional models in the traditional regression task, while the GNN-based model performs much better when combined with the causal weighting. As for uplift modeling estimation, as shown in Figure 6, causal weighting combined architecture has a much more apparent effect. Before combining causal weighting information, GNN based model is slightly

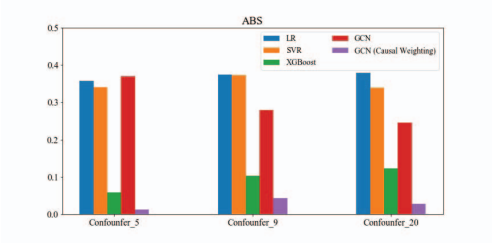

Fig. 6. Absolute error of ITE of base methods and ours with the numbers of confounders are 5, 9 and 20, respectively.

TABLE I
AUUC FOR UPLIFTING EVALUATION AND MSE FOR Y-PREDICTION OF BASELINE METHODS AND PROPOSED METHOD

Model	AUUC	MSE
LR	0.4980	0.0026
SVR	0.5475	0.0037
XGBoost	0.8756	0.0025
GCN	0.5443	0.0028
GCN (Causal Weighting)	0.8807	5e-06

better than LR and SVR in the estimation of uplift value but worse than xgboost while achieving a very accurate result when adopting the causal weighting combined architecture. Another result is that GNN based model can have a much more stable performance when the number of confounders increases, which means that it can have a much more robust performance when facing more complex situations.

2) Real-world dataset: In a real-world dataset, the actual causal effect of treatment remains unknown, leading to the abovementioned indicators being inapplicable. As a result, Area Under Uplift Curve (AUUC) is adopted to measure the performance of an uplifting model on the real-world dataset. AUUC can be calculated as follows:

$$AUUC(f) = \int_0^1 V(f,x)dx \approx \sum_{k=1}^{n} V(f,k) \tag{23}$$

$$whereV(f,k) = \frac{1}{|T|}\sum_{i\in f(\mathcal{D},k)} y^1_{i\,[t_i=1]} - \frac{1}{|C|}\sum_{j\in f(\mathcal{D},k)} y^1_{j\,[t_j=0]} \tag{24}$$

Here $f(\mathcal{D},k)$ can be the k first samples of the dataset when ordered by the prediction of the model f, $|T|$ is the number of samples in the treatment group(t=1), and $|C|$ is the number of samples in the control group(t=0).

For certain causal relationships, the higher the AUUC is, the better the uplifting model performs. The AUUC of the baseline models and ours are listed in Table I.

Table I shows that when estimating the uplift value in the real-world dataset, although origin GNN has a similar performance with LR and SVR, it has a better performance than XGBoost when with the causal weighting combined architecture.

V. Conclusion and Future Work

In this work, we investigated how to describe causal information in uplift modeling (add conditional average treatment effect (CATE) and build an adjacency matrix using Bayesian network structure learning). In addition, we addressed how to incorporate this causal information into uplift estimations by proposing a framework for uplift modeling that is based on graph neural networks.

Experiments on simulated and real-world datasets reveal that while the origin graph convolutional neural network performs comparably to conventional approaches when directly predicting uplift values, when paired with causal neighbourhood features and causal representation information, it demonstrates exceptional performance in both the prediction job and the uplift estimation task of the target, owing to the GCN's excellent neighbourhood learning features.

It is worthwhile to investigate more methods of characterising causal knowledge in the future. Weighted adjacent matrices might be seen as a means of guiding graph convolutional neural networks to provide accurate data. Alternatively, it is equally intriguing to investigate the size of the receptive domain of neighbourhood features. A wider receptive domain denotes more information, which might aid us in enhancing the performance of this job in downstream prediction.

References

[1] E. Diemert, A. Betlei, C. Renaudin, and M.-R. Amini, "A large scale benchmark for uplift modeling," in *KDD*, 2018.

[2] M. L. Whittal, S. R. Woody, P. D. McLean, S. Rachman, and M. Robichaud, "Treatment of obsessions: A randomized controlled trial," *Behaviour research and therapy*, vol. 48, no. 4, pp. 295–303, 2010.

[3] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[4] J.-F. Ton, D. Sejdinovic, and K. Fukumizu, "Meta learning for causal direction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 9897–9905, 2021.

[5] H. Chen, T. Harinen, J.-Y. Lee, M. Yung, and Z. Zhao, "Causalml: Python package for causal machine learning," *arXiv preprint arXiv:2002.11631*, 2020.

[6] M. Caliendo and S. Kopeinig, "Some practical guidance for the implementation of propensity score matching," *Journal of economic surveys*, vol. 22, no. 1, pp. 31–72, 2008.

[7] F. Li, K. L. Morgan, and A. M. Zaslavsky, "Balancing covariates via propensity score weighting," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 390–400, 2018.

[8] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International conference on machine learning*, pp. 10–18, PMLR, 2013.

[9] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: a comprehensive review," *Computational Social Networks*, vol. 6, no. 1, pp. 1–23, 2019.

[10] H. Wang, X. Ye, Z. Zhang, and Y. Wang, "Multihead causal distilling weighting is all you need for uplift modeling," in *2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, pp. 59–65, IEEE, 2022.

[11] N. C. Chesnaye, V. S. Stel, G. Tripepi, F. W. Dekker, E. L. Fu, C. Zoccali, and K. J. Jager, "An introduction to inverse probability of treatment weighting in observational research," *Clinical Kidney Journal*, vol. 15, no. 1, pp. 14–20, 2022.

[12] H. J, "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," in *Political analysis*, pp. 20(1): 25–46, 2012.

[13] S. Athey, G. W. Imbens, and S. Wager, "Approximate residual balancing," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 80, no. 4, pp. 597–623, 2018.

[14] K. A. Bollen, "An alternative two stage least squares (2sls) estimator for latent variable equations," *Psychometrika*, vol. 61, no. 1, pp. 109–121, 1996.

[15] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," *Advances in neural information processing systems*, vol. 31, 2018.

[16] F. P. Tso, L. Cui, L. Zhang, W. Jia, D. Yao, J. Teng, and D. Xuan, "Dragonnet: a robust mobile internet service system for long-distance trains," *IEEE transactions on mobile computing*, vol. 12, no. 11, pp. 2206–2218, 2013.

[17] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," *Advances in neural information processing systems*, vol. 30, 2017.

[18] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[19] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey, "Double/debiased/neyman machine learning of treatment effects," *American Economic Review*, vol. 107, no. 5, pp. 261–265, 2017.

[20] A. A. Neath and J. E. Cavanaugh, "The bayesian information criterion: background, derivation, and applications," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 2, pp. 199–203, 2012.

[21] B. Selman and C. P. Gomes, "Hill-climbing search," *Encyclopedia of cognitive science*, vol. 81, p. 82, 2006.

[22] J. H. Friedman, E. Grosse, and W. Stuetzle, "Multidimensional additive spline approximation," *SIAM Journal on Scientific and Statistical Computing*, vol. 4, no. 2, pp. 291–301, 1983.