Multihead Causal Distilling Weighting Is All You Need for Uplift Modeling

Haowen Wang*
Business Algorithm Department
Alipay, AntGroup
Shanghai, China
wanghaowen.whw@antgroup.com

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

Zhiyi Zhang
Dinosaur Tech
Shanghai, China
emma0302@pku.edu.cn

Yikang Wang
Centre for Advanced Spatial Analysis
University College London
London, UK
yikang.wang.21@ucl.ac.uk

2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS) | 978-1-6654-9144-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/HDIS56859.2022.9991731

***Abstract*—Uplift modeling is a branch of predictive modeling technology, which is usually used to analyze marketing, advertising and product personalization experiments. In this kind of application scenario, we usually do a large number of random experiments to assist decision-making. And thus there is a great need for us to control the number of features as well as tell the most important features in order for the next steps. What's more, in this kind of scenario, the interpretability of features also becomes very important, which means that we should not only pursue the accuracy of prediction, but also minimize the difficulty of controlling the features. However, most existing studies only focus on the accuracy of model prediction, but ignore the cost of controlling and observing too many variables as well as the interpretability of features in practical application. In order to better solve this problem, we introduce a multi head weight calculation method based on causal inference. Instead of selecting features based on the result of machine learning, we manage to select features from the source with causal inference, a total different method from traditional machine learning methods. It can be viewed as a method to solve the problem of overfitting. In our experiment, we use the data from different industries and use different numbers of selected features to evaluate the effectiveness of the proposed feature selection method. The results show that our algorithm significantly improves the performance compared with the feature selection method in standard machine learning theory.**

***Index Terms*—Uplift modeling, Causal inference, Feature selection, Multihead, Causal Distilling**

I. Introduction

With the deepening of the digitization of the world, the idea of random test, or A/B test, has gradually penetrated into various marketing fields, such as advertising, recommendation system, customer service, promotion and product design. How to use the observed marketing data to adjust the marketing strategy to maximize the revenue has become a common problem in more and more industries. For a long time, uplifting model [1] has been considered as a technical designed method to estimate the treatment effect, and has been used in the actual optimization process in many fields.

Uplift modeling is different from traditional machine learning and deep learning. It serves more specific marketing processes, which are mostly the extension of decision-making. As a result, we can not only care about the accuracy of the model, but also put forward higher requirements for how to select more informative, predictive and relatively interpretable features. However, the past research and discussion on marketing estimation and uplift modeling mostly focused on how to design the model structure to estimate more accurate ATE and ITE values [2], [3], but ignored the important role of feature selection in the actual marketing process in such practical application scenarios, that is to say, an evaluation model that needs to control 100 features may be meaningless to the actual marketing process, because the cost of feature control is unaffordable for the actual marketing process.

Generally speaking, in order to reduce the observation and operation costs of enterprises, we hope to select as few features as possible on the basis of ensuring the estimation quality. Another thing to note is that using all available features will make the calculation of the model inefficacy [4], and it is easier to lead to over fitting [5], reduce the interpretability of the model and reduce the intervenability of the marketing process. Consequently, feature selection plays an important role in marketing data mining and uplift modeling estimation, which can make full use of rich feature information and reduce correlation loss.

However, although feature selection is of great significance to the actual marketing process and uplift modeling, it has rarely been discussed in the past work. At the same time, feature selection in machine learning theory has been well studied. However, we will use two groups of experiments to show that the existing machine learning feature selection methods are ineffectual for actual marketing data mining and uplift modeling. Therefore, it is necessary to carry out the research and exploration of interpretable feature selection for the marketing process with practical significance, especially considering the uplift modeling based on it.

One challenge we encounter is that in the actual marketing process, the traditional marketing data is often observation data and the number of samples is small. The sample imbalance and small sample problems make the machine learning algorithm based on statistics easier to over fit. Another practical problem is that in the actual business environment, especially in the scenario concerned by uplift modeling represented by marketing, the market is constantly changing, which means that the statistical two-way correlation between features is easy to fail.

We now describe the contribution of our paper on method-

* Corresponding Author.

978-1-6654-9144-0/22/$31.00 ©2022 IEEE

ology and empirical evaluation perspectives. We propose a feature importance weight calculation method (MCDW)based on causal inference and knowledge distillation to address. Firstly, we use the causality diagram model framework to solve the knowledge learning problem of small samples. Secondly, the framework based on causality diagram model is more extendable, which can make better use of a priori expert knowledge to mine the causality between features and target variables. In addition, the idea of knowledge distillation can be well compatible with our computational framework, which improves the generalization performance of feature selection and effectively reduces overfitting. Finally, our feature importance weight calculation framework is multi-head, which means that its calculation can achieve efficient parallelism, which makes it possible for its large-scale commercial use.

In this paper, we focus on two marketing problems, the feature selection problem of classification and regression problem. In the first example, the outcome variable is a continuous revenue and in the second example, the outcome variable is a categorical customer conversion value. It is worth mentioning that through knowledge distillation, we can transform the classification problem into a regression problem to a certain extent.

II. Related Work

A. Uplift Modeling

Uplift model can be regarded as a method to estimate the incremental feedback value of heterogeneous treatment at the user level. Usually, this method can be combined with machine learning. According to the most commonly used Potential Outcome Framework(POF) [6] introduced by Neyman Rubin, the individual treatment effect can be expressed as:

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

where $Y_i(1)$ and $Y_i(0)$ represents the result of the outcome variable under treatment condition and control condition respectively for individule i.

Considering that the individual effect of treatment will vary from individuals and the high cost of marketing experiments in the industry, we pay more attention to the effect of treatment on user groups. Therefore, the conditional average treatment effect (CATE) is introduced [7]:

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

whereX_i is the feature vector for user i.

Since the calculation of treatment effect by cate is based on the vector represented by the observation user's feature, it will have higher flexibility. Different treatment conditions can be selected. We can also calculate the individual treatment effect by adjusting the feature vector as needed.

The most direct uplift modeling method is separate model approach (SMA) [8], which uses the separated model to predict the corresponding response of users in each category [9]. This process enables the regression and classification algorithms in supervised learning to be deployed directly. However, although the separate model approach is a simple and effective criterion, it may perform not well in a real business environment. This is because SMA tends to predict the response rather than find the uplift signal, which is illustrated before in.

Considering the shortcomings of SMA, researchers have proposed a lot of algorithms aiming at modeling the uplift process directly. A logistic regression method is first proposed to estimate the explicit relationship between features and treatment [10]. Then the application of k-nearest neighbors (KNN) method is introduced [11], [12]. KNN can be used to find sample objects with similar characteristics, and then observe the action response on these similar samples [9]. The ability of support vector machines to find and partition hyperplanes was later used to estimate the different effects of treatment [13]. Tree based modeling of uplift signal can also avoid this weakness, many past facts have proved that this evolutionary algorithm based on selecting the attribute with the maximum splitting criterion and maximizing the entropy after splitting has a good effect on estimating the response of action and predicting the target variable directly [12], [14], [15]. This is also the main reason why we use the tree model as the benchmark method as the downstream prediction and evaluation process in this paper.

B. Machine Learning based Feature Selection Methods

Feature selection is the process of selecting a subset of relevant features for model building, which simplifies models to make them easier to interpret by researchers/users [16]. For the design of feature selection, there has been a lot of studies by scholars. There are roughly three types of common feature selection methods, filter, embedding and wrapper [17]. The filtering method will filter the data set before training the learner, and the feature selection process has nothing to do with the subsequent learner; the embedded method integrates the feature selection process with the learner training process, and automatically performs feature selection during the learner training process [18] [19], one typical method is feature importance calculation based on XGBoost [20]. The wrapping method directly uses the performance of the final learner as the evaluation criterion for feature subsets, and it is usually proved that the classification accuracy of searching feature subsets is better than the former two [21], recursive feature elimination [22] is considered as an efficient framework of this type. In the experiment of this paper, we use XGBoost and RFE framework to calculate the feature weight as the comparison of our MCDW method.

1) XGBoost: XGBoost is an efficient system implementation of the idea of gradient boosting, and its base learner can be either a linear classifier or a tree. This paper uses the characteristics of its tree model as the basis to quantify the importance of each feature for feature selection.

The importance metric is a measure to evaluate the importance of each feature in the feature set to which it belongs [23]. XGBoost uses the number of feature splits FScore, the average feature gain value AverageGain and the feature aver-

age coverage AverageCover as the basis for its decision tree construction, so as to accurately complete the classification task [24].

For the above three importance metrics, we have

$$\mathrm{FScore} = |X| \tag{3}$$

$$\mathrm{AverageGain} = \frac{\sum \mathrm{Gain}_X}{\mathrm{FScore}} \tag{4}$$

$$\mathrm{AverageCover} = \frac{\sum \mathrm{Cover}_X}{\mathrm{FScore}} \tag{5}$$

Among them, X is the set of the required feature classification to leaf nodes; Gain is the node gain value obtained by formula (4) for each leaf node in X at the time of segmentation; Cover is the number of samples in X that fall on each node.

2) SVM Recursive Feature Elimination (SVM RFE): SVM-RFE is a method of feature ranking with recursive feature elimation [25]. Given an external estimator(SVM) that assigns weights to features (in our experiment, the square of the weight vector of dimension length(s)), the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller feature sets. Firstly, the estimator is trained on the initial feature set, and the importance of each feature is obtained through any specific attribute or callable attribute. Then, the least important features are deleted from the current feature set. The process repeats recursively on the pruning set until the required number of features is finally reached.

The resulting function of an input vector $\mathbf{x}$ is:

$$D(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{6}$$

with

$$\mathbf{w} = \sum_k \alpha_k y_k \mathbf{x}_k \quad \text{and} \quad b = \langle y_k - \mathbf{w} \cdot \mathbf{x}_k \rangle \tag{7}$$

The weight vector $\mathbf{w}$ is a linear combination of training patterns. The ranking criteria is $c_i = (w_i)^2$, for all i. In RFE, we remove the feature with smallest ranking criterion.

C. Causal inference

It is generally believed that the framework of causal inference originates from D.B. Rubin's Rubin casal model (RCM) [26], which is a framework based on potential outcome. Because it envisages situations contrary to observations, it is a kind of counterfactual causality. It also shows the status of random test as the golden criterion for causal inference. It is worth mentioning that the influence of this framework is further expanded due to the use of difference in difference (DID) [27] in randomized trials. Since then, the methods of exact matching and proportion score matching (PSM) [28] based on random trials have been applied, but causal inference still faces the challenges of the number of users and high-dimensional characteristics.

Then, based on Rubin's basic law of counterfactual, Judea pearl proposed structural causal model (SCM) [29]. Combined with Bayesian D-seperation, he proposed the concept of causal reasoning engine, that is, inputting assumptions and data with graphical model as the carrier, and outputting estimate, estimate and fit indexes.

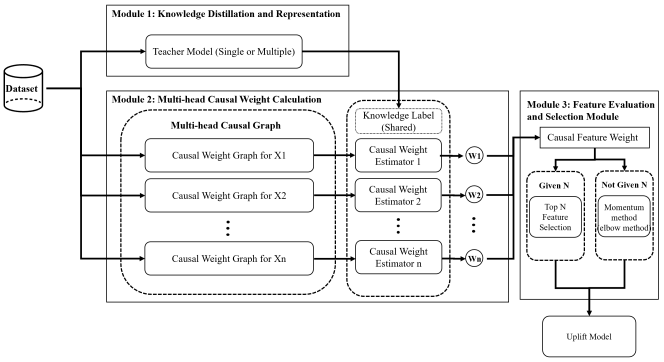

Fig. 1. Overall framework of Multihead Causal Distilling Weighting(MCDW) method

Under this framework, more statistical machine learning theories can be applied. The tree model represented by BART [30] and causal forest [31] produces the greatest difference in processing effect between different leaves, but still gives us an accurate estimation of processing effect. In addition, the introduction of double robust method [32] successfully reduces the estimation deviation. The meta learning method [33] optimizes the small sample and the large sample proportion deviation with or without treatment. In addition, adversarial learning [34] and representation learning [35] are also applied in this new framework to further improve the estimation process.

D. Knowledge Distilling

Knowledge distillation [36] is a teacher-student training structure. Generally speaking, it uses the trained teacher model to provide knowledge. The student model obtains the teacher's knowledge through distillation training. This process can transfer the knowledge of complex teacher model to simple student model. Later, due to different application scenarios, model compression and model enhancement were developed based on knowledge distillation.

The work of knowledge distillation related to this paper focuses on the representation and learning of knowledge and the enhancement of model. In the multi classification problem, knowledge is represented in the form of soft-label, and the concept of distillation temperature is introduced. In the regression problem and binary classification problem, the representation of knowledge becomes simpler due to the collapse of dimension. In the subsequent research, Fitnets first proposed the method of learning knowledge through intermediate feature layers, which has better generalization. Combined with the idea of integrated learning, multiple teacher networks has been proved to be better than single teacher in some tasks.

III. Methodology

Fig. 1 shows the overall framework of Multihead Causal Distilling Weighting(MCDW) method. It includes three modules, including knowledge distillation and representation mod-

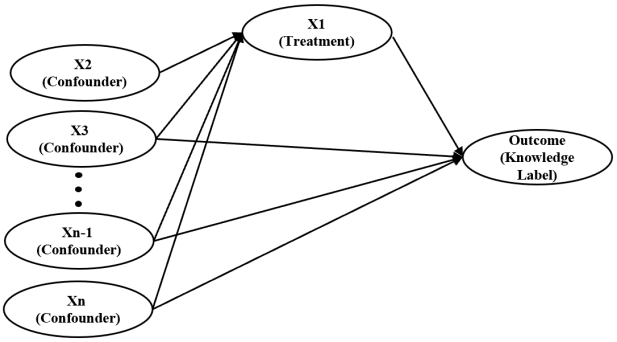

Fig. 2. Causal Weight Graph mode 1 for Feature X_1

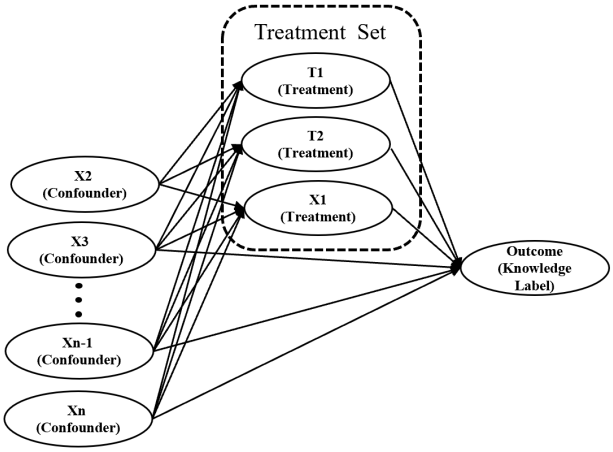

Fig. 3. Causal Weight Graph mode 2 for Feature X_1

ule, multi-head causal weight calculation module and feature evaluation and selection module.

A. Knowledge Distillation and Representation Module

The knowledge distillation and representation module can use the teacher network to pre-train the data, and store the learned knowledge in the output, that is, the knowledge label. It is worth mentioning that we can support multiple distillation architectures of single teacher and multi teacher.

In the problem of discrete targets, soft label contains more information than hard label, because it retains the information of different categories of relative probability. Similarly, in the problem of continuous goals, previous studies have also proved [37] that the distilled goal value is more conducive to information retention and knowledge learning and training.

B. Multi-head Causal Weight Calculation Module

The multi head causal weight calculation module is established based on the neural network of causality diagram. Firstly, we establish n causal weight graph in parallel, one causal weight graph for X_1 is shown in Fig. 2, we set X_1 as treatment while other features as confounders.

If we have a priori expert knowledge on feature selection, we can control the graph network and adjust the features by setting the treatment set, as shown in Fig. 3. Firstly, we fix the known treatment, and add the remaining features to the treatment set, and the other features are regarded as confounders. In this way, we can combine the prior knowledge of feature selection into graph network.

In this way of building the causal graph network, we ensure that the status of features is symmetrical in the n causal weight graph models. Then on the basis of each graph model, we estimate the cause and effect value of each feature as the target when intervening. The causal value corresponding to each feature is saved as the feature weight based on causality. Here, the training target is the knowledge label obtained in the first module. Since the graph models for estimating the causal effects of each feature are independent of each other, it is possible for us to set up multi-head process with parallel computing. Considering the adjustability of causality diagram model, we support embedding a priori-knowledge into the graph model in the form of network structure for learning.

For the estimation process of CATE of every feature, we use the method of a causal forest combined with double machine learning based residualization of the treatment and outcome variable. Here we define the absolute value of CATE as causal weight because we are more concerned about the extent to which the feature exerts a causal effect on the target than the direction (positive or negatieve) of causal effect it exerts on the target. You can also use other methods based on double machine learning or even meta learning methods. Our calculation framework has good sustainability for these estimation methods.

C. Feature Evaluation and Selection Module

The feature evaluation and selection module includes two methods for selecting the number of features. The first is given the number of selected features. We will rank the five most influential features according to the weight of the features. The second is the momentum based selection method. When the number of features is not specified, we select the feature with the fastest decline in feature weight as the dividing line, and select the feature before the turning point, that is, the elbow method.

IV. Experiment and Discussion

A. Dataset

The dataset [38] contains data of 2,000 customers of a startup company that sells software. Each customer has eight features including four Boolean values (having global offices, being a large consumer, being a small medium corporation, is commercial) and four continuous variables (IT spent, employee count, PC count, and size), two interventions showing incentive given to customers (technical support and discount), and one outcome showing the amount of products purchased by the customer within one year of the incentives (revenue). The correlations of these variables are shown in Fig. 4.

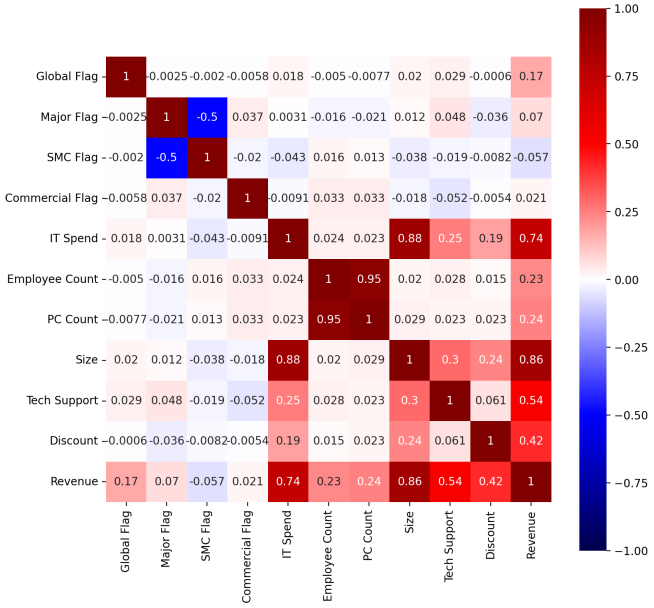

Fig. 4. The correlation of the variables in the dataset.

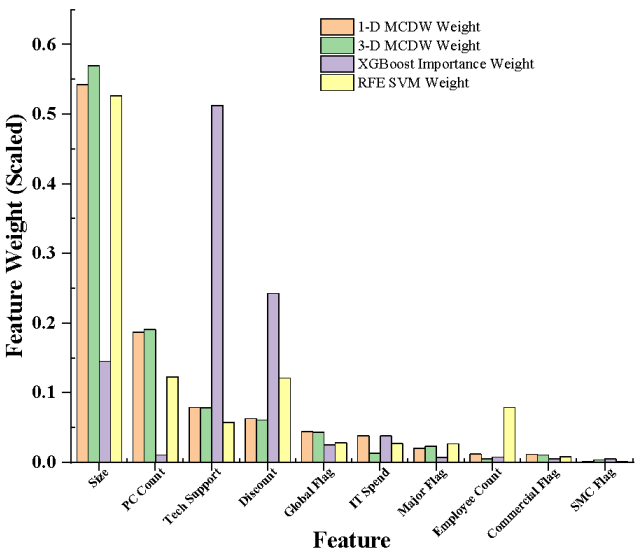

Fig. 5. Output of Three Weight Calculations.

B. Experiment Design

We designed experiments to verify and evaluate the performance of our model. The data set we use here is the user marketing data set introduced in the first subsection. Our task here is to evaluate the rationality of feature selection through the prediction of revenue. We choose lightGBM as the prediction model for feature selected from different method, which is a more efficient and more accurate algorithm for GBDT framework. In terms of feature selection, we choose two methods as the baseline, which are the feature selection method based on xgboost feature importance and the feature selection method based on recursive feature optimization (RFE). The estimator in RFE selects the support vector machine (SVM) based on the idea of hyperplane division. In this experiment, we select xgboost single model for knowledge distillation in the first module and we select using a given number of features mode in the third module.

C. Result and Discussion

Fig. 5 shows the calculation results of feature weight under the default conditions (1-D treatment estimation) and a priori knowledge embedding (3-D treatment estimation) under the framework of multihead causal distilling weight (MCDW) method as well as the results of traditional xgboost feature importance calculation and RFE-SVM feature weight. The results of 1-D and 3-D in MCDW show that the weight calculation method based on causal graph has good robustness and can make a more stable estimation of the causal weight of feature to y.

There are obvious differences between the calculation methods of MCDW and xgboost. For example, for feature *PC Count*, the weight of xgboost's feature estimation method is low, but it is very important in MCDW method. The same thing happens in feature *Size*, but it has the opposite performance in *Tech Support* and *Discount*. Another common feature selection method is Recursive Feature Elimination (RFE). This method deletes the features that contribute less to the accuracy rate from the result level through overall optimization. There is still a large gap between its result and MCDW's calculation, especially for the weight calculation of feature *employee count*.

To some extent, 1-D estimation in MCDW framework is similar to Shapley value [39] in cooperative game theory. However, although they all retain the similar condition of single variable, MCDW framework focuses on inferring causality from causal graph, which is essentially the embodiment of causality, while Shapley value is estimated from the perspective of removing a feature from the framework of correlation, thus the quantified value of shapley is biased, and the specific value has no reference significance. Although it can reflect a certain causality from the perspective of processing simple information (whether there is a feature or not), in complex relationships (such as nonlinear relationships), its estimation and interpretation ability is limited. As Fig. 6 shows, they differ in the analysis of characteristics *Global Flag*.

After we calculate the feature importance weight, we can select the features according to the number of features we need, Here we select the number of features range from 1 to 5 to analyze the prediction effect and evaluate the feature weight. Since the feature importance ranking of 1-D MCDW estimation and 3-D MCDW estimation is the same in 1-5, we combined their results together as MCDW Weight function. Table.I compares the results when we use different feature weights.

In general, the feature selection method based on MCDW calculation weight performs more stable and better in different scenes with a given number of features, which is more obvious when selecting fewer features comparing with XGBoost

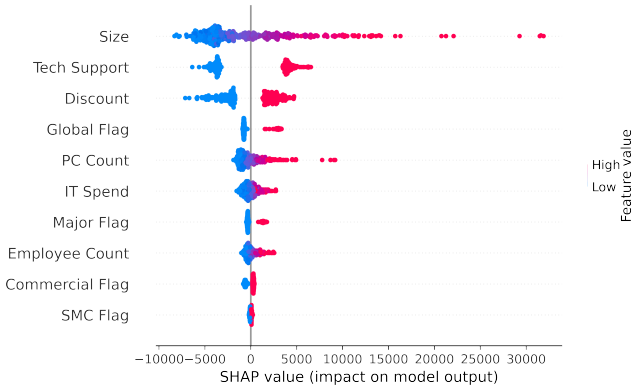

Fig. 6. Shapley Value Analysis Result with the Dataset.

TABLE I
PERFORMANCE WITH DIFFERENT FEATURE SELECTION FRAMEWORK

Feature Number	MCDW Weight		XGBoost Importance Weight		RFE SVM Weight	
	MSE Loss	R^2	MSE Loss	R^2	MSE Loss	R^2
1	0.005209	0.696958	0.012556	0.269478	0.005424	0.684448
2	0.004609	0.731863	0.009755	0.432459	0.004680	0.727723
3	0.003413	0.801407	0.003007	0.825038	0.003560	0.792899
4	0.001940	0.887110	0.002921	0.830050	0.002211	0.871369
5	0.002112	0.877152	0.002639	0.846461	0.001966	0.885609

selecting function. Although the performance of SVM method based on RFE framework is similar to that of MCDW, it lacks the explanatory ability of feature weight. In fact, if we restore the original symbol of MCDW weight, we can even get how these features affect the results (positive or negative).

The reason why MCDW method can achieve better results comes from its redefinition of feature importance weight. We introduce causal effect into the definition of feature importance weight for the first time. We believe that the causal relationship between features is always more stable than correlation, because there is an interpretable logical chain of causality, which is less likely to fail. Although in many cases (if there are enough features selected without considering the cost), the correlation can achieve the results similar to the causality analysis. For the fields that have an impact on the reality, such as the marketing field, considering the cost, when the number of features selected is limited, the features selected based on the causality framework undoubtedly have robustness and universality to more changeable market situations.

V. CONCLUSION AND FUTURE WORK

This paper proposes a feature importance weight calculation method: Multihead Causal Distilling Weighting(MCDW) based on graph network causal reasoning, and verifies the actual marketing experimental dataset. We compare the xgboost feature importance calculation method commonly used in the industry and the feature selecting method based on RFE-SVM, which has achieved more stable and excellent performance in the task of a given number of features. Although the performance is similar to that of RFE-SVM, the feature weight of MCDW is more interpretable, and its positive and negative causality can be used as an important reference for subsequent actual marketing process intervention. We have proved that MCDW method works well in feature modeling and screening of actual marketing process, but its role is still limited for application scenarios with more features and a larger amount of data. Future work is emphasized in the following aspects: (1) Embedding more complex graph network to deal with application scenarios with larger amount of data and more features. (2) Extending MCDW method from the field of uplift modeling with cost constraints to more machine learning and even deep learning tasks. (3) Making more efficient use of the feature importance weight estimated by MCDW method and make more use of the amount of information obtained in feature analysis in the model.

REFERENCES

[1] Robin Marco Gubela, Stefan Lessmann, Johannes Haupt, Annika Baumann, Tillmann Radmer, and Fabian Gebert. Revenue uplift modeling. In *Machine Learning for Marketing Decision Support*. 2017.

[2] Olav Vestøl, Jonas Ågren, Holger Steffen, Halfdan Kierulf, and Lev Tarasov. Nkg2016lu: a new land uplift model for fennoscandia and the baltic region. *Journal of Geodesy*, 93(9):1759–1779, 2019.

[3] Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, volume 46, pages 79–95, 2012.

[4] Jacob Sheinvald, Byron Dom, and Wayne Niblack. A modeling approach to feature selection. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume 1, pages 535–539. IEEE, 1990.

[5] Xue Ying. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing, 2019.

[6] Janet M Box-Steffensmeier, Henry E Brady, David Collier, et al. *The Oxford handbook of political methodology*, volume 10. Oxford Handbooks of Political, 2008.

[7] Dana Burde and Leigh L Linden. Bringing education to afghan girls: A randomized controlled trial of village-based schools. *American Economic Journal: Applied Economics*, 5(3):27–40, 2013.

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

[9] Chenchen Li, Xiang Yan, Xiaotie Deng, Yuan Qi, Wei Chu, Le Song, Junlong Qiao, Jianshan He, and Junwu Xiong. Reinforcement learning for uplift modeling. *arXiv preprint arXiv:1811.10158*, 2018.

[10] Victor SY Lo. The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, 4(2):78–86, 2002.

[11] Farrokh Alemi, Harold Erdman, Igor Griva, and Charles H Evans. Improved statistical methods are needed to advance personalized medicine. *The open translational medicine journal*, 1:16, 2009.

[12] Xiaogang Su, Joseph Kang, Juanjuan Fan, Richard A Levine, and Xin Yan. Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research*, 13:2955, 2012.

[13] Łukasz Zaniewicz and Szymon Jaroszewicz. Support vector machines for uplift modeling. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 131–138. IEEE, 2013.

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

[15] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327, 2012.

[16] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[17] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[18] Zhi-Hua Zhou. *Machine learning*. Springer Nature, 2021.

[19] Dai Chunlei and Bai Jing. Application of multi-label classification algorithm based on embedded feature extraction in financial management optimization system. In *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1384–1387. IEEE, 2021.

[20] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[21] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502, 2005.

[22] Xiaohui Lin, Fufang Yang, Lina Zhou, Peiyuan Yin, Hongwei Kong, Wenbin Xing, Xin Lu, Lewen Jia, Quancai Wang, and Guowang Xu. A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. *Journal of chromatography B*, 910:149–155, 2012.

[23] Huiting Zheng, Jiabin Yuan, and Long Chen. Short-term load forecasting using emd-lstm neural networks with a xgboost algorithm for feature importance evaluation. *Energies*, 10(8):1168, 2017.

[24] Mingfei Hu, Xinyi Hu, Zhenzhou Deng, and Bing Tu. Fault diagnosis of tennessee eastman process with xgb-avssa-kelm algorithm. *Energies*, 15(9):3198, 2022.

[25] Jason Weston and Isabelle Guyon. Support vector machine—recursive feature elimination (svm-rfe), January 10 2012. US Patent 8,095,483.

[26] Guido W Imbens and Donald B Rubin. Rubin causal model. In *Microeconometrics*, pages 229–241. Springer, 2010.

[27] Stephen G Donald and Kevin Lang. Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, 89(2):221–233, 2007.

[28] Peter C Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine*, 27(12):2037–2049, 2008.

[29] Judea Pearl. *Causality*. Cambridge university press, 2009.

[30] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[31] Jonathan Davis and Sara B Heller. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–50, 2017.

[32] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[33] Laura Hennefield, Hyesung G Hwang, Sara J Weston, and Daniel J Povinelli. Meta-analytic techniques reveal that corvid causal reasoning in the aesop's fable paradigm is driven by trial-and-error learning. *Animal cognition*, 21(6):735–748, 2018.

[34] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017.

[35] David Heckerman. A bayesian approach to learning causal networks. *arXiv preprint arXiv:1302.4958*, 2013.

[36] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[37] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

[38] Haowen Wang. Software Usage Promotion Campaign Uplift Modeling. Type: dataset.

[39] Eyal Winter. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054, 2002.