Causal Inference in Online Advertising

Can Chen Jintong Luo Xiaotian Shao Jingyi Zhou
Department of Data Science
UC San Diego
La Jolla, CA 92093
cac024@ucsd.edu, jil386@ucsd.edu, x5shao@ucsd.edu, jiz238@ucsd.edu

Abstract

Advertising can yield significant profits, but targeting the wrong audience can have adverse effects. Therefore, identifying the right targets is crucial, and uplift modeling is a popular approach to address this challenge. In this project, we utilized two methods: meta-learning (T-learner) and direct modeling (causal graph based on graph neural networks) to model the uplift effect on the Criteo dataset. Our results demonstrate that high performance (AUUC over 0.85) can be achieved using simple meta-learning methods. Additionally, we observed the instability of causal discovery when using graph neural networks.

1 Introduction

1.1 Challenges in online advertising

Online learning is a crucial component of marketing strategies. In online advertising, accurately assessing campaign effects through causal inference is crucial for refining strategies and optimizing budgets. Measuring the advertising lift effect, or the increment in user conversion, is a key task in digital brand marketing. Typically, this involves comparing conversions between a control group, which does not receive ads, and a treatment group, which does. However, traditional observational studies lack precision due to non-homogeneous groups Lewis et al. (2011); Varian (2016). While randomized controlled trials (RCTs) offer a solution by creating homogeneous groups, real-world application faces challenges like exposure bias, affecting the data quality and measurement accuracy of uplift effects Gordon et al. (2019); Diemert et al. (2018); Liu and Chamberlain (2018).

Uplift modeling aims to estimate the uplift effect, also known as the individual treatment effect (ITE) Devriendt et al. (2018). When resources are limited, this method can enable rapid and personalized decision making, which is more efficient than RCTs. By evaluating the impact of an intervention on different groups, we can determine which customers are most likely to change their behavior as a result of the intervention. Existing two-model approaches typically model two separate potential outcomes for the treatment and control groups, and then estimate the upward effect based on the difference between the outputs of the two modelsHansotia and Rukstales (2002); Hill (2011); Cai et al. (2011). In addition, single-model approaches to upward modeling can combine the potential outcome models for the treatment and control groupsHansotia and Rukstales (2002); Hill (2011); Cai et al. (2011).

1.2 Related Work

Uplift modeling aims to identify specific subgroups that benefit most from a predetermined action, it has received limited attention in the literature, despite the fact that this approach is widely used in practice The concept was first discussed in depth by Radcliffe and Surry (1999), highlighting its utility across various use cases. Further discussions and applications of uplift modeling are elaborated in sub-

Preprint. Under review.

sequent works by Hansotia and Rukstales (2002) and Radcliffe and Surry (2011). Related techniques such as A/B testing and action rule discovery, discussed by Kohavi et al. (2009) and Adomavicius and Tuzhilin (1997) respectively, focus on overall effectiveness or specific action outcomes, differing from uplift's targeted subgroup approach. In terms of technical methods, ensemble approaches like bagging, boosting, and Random Forests are predominant, with Breiman's works (Breiman (1996), Breiman (2001)) being central, alongside other methods such as Extremely Randomized Trees and Random Decision Trees which inject randomness into model learning to promote diversity within ensembles.

1.3 Objectives

Uplift models focus on evaluating the net effect of a treatment by identifying the differential response of customers to communication. According to Vanguarde (2022), these models categorize customers into four groups based on their actions (Y) and treatment status (W):

1. **Will Take Action**: Individuals who will complete the action regardless of treatment ($Y = 1, W = 1$ and $Y = 1, W = 0$).
2. **Persuadable:** Individuals who only respond positively to the action when treated ($Y = 1, W = 1$ and $Y = 0, W = 0$).
3. **Do Not Disturb:** Individuals who react negatively when treated ($Y = 1, W = 0$ and $Y = 0, W = 1$).
4. **Never Will Respond:** Individuals who are indifferent to treatment ($Y = 0, W = 1$ and $Y = 0, W = 0$).

This segmentation allows for the targeted application of treatments to optimize responses.

We implemented two methods to analyze the uplift effect on Criteo Uplift Prediction Dataset, including the T-learner method and a direct modeling method using the integration of Graph Neural Networks with causal knowledge. We compare the results with two metrics (AUUC,AUQC) and we find some limitations that can be refined in the future.

2 Methodology

2.1 Problem formulation

The uplift modeling problem can be formalized as follows, based on Kawanaka and Moriwaki (2019). We consider a set of N individuals indexed by i. Let Y_i^T denote the potential outcome for individual i when they receive the treatment, and Y_i^C denote the potential outcome when they do not receive the treatment (Control). In the context of online advertising, the treatment is sending an advertisement, and the outcome of interest is whether the individual converts (i.e., buys the product). The individual treatment effect (ITE) τ_i is defined as the difference in outcomes for individual i between receiving the treatment and not receiving it:

$$\tau_i = Y_i^T - Y_i^C \tag{1}$$

However, we can only observe whether an individual converts or not, and an individual may convert given the advertisement this time and not convert the next time. Therefore, we are more interested in the average likelihood of conversion for this individual. Moreover, since the ITE varies from individual to individual and the cost increases with more individuals in experiments, we are often more interested in the conditional average treatment effect (CATE), which is the average treatment effect on subgroups:

$$\tau_i = E[Y_i^T|X_i] - E[Y_i^C|X_i] \tag{2}$$

where X_i is individual i's feature vector.

If random controlled trails are satisfied, we have the conditional ignorance assumption satisfied: the potential outcomes $\{Y_i^T, Y_i^C\}$ are independent of their treatment assignment W_i conditioned on their feature vector X_i:

$$Y_i^T, Y_i^C \perp\!\!\!\perp W_i \mid X_i \tag{3}$$

Using this assumption, we can estimate CATE even though we cannot observe both Y_i^T and Y_i^C simultaneously:

$$\tau_i = E[Y_i^T|X_i, W_i = 1] - E[Y_i^C|X_i, W_i = 0] \tag{4}$$

Moreover, we can simplify the individual i's actual observed outcome as:

$$y_i^{obs} = W_i Y_i^T + (1 - W_i) Y_i^C \tag{5}$$

Here, $W_i \in \{1, 0\}$ is a binary indicator taking on the value 1 if i receives the advertisement and 0 if not. The unobserved conversion probability is estimated using regression methods on users' feature vectors X_i.

The purpose of uplift modeling is to find a function f such that:

$$\tau_i = f(X_i). \tag{6}$$

In this project, we use meta-learning methods and direct modeling methods based on Graph Neural Networks (GNNs) to estimate the treatment effects.

2.2 Dataset

The Criteo dataset is a pivotal resources for causal inference and uplift modeling. The dataset , released by Criteo Labs, consists of 13M rows, featuring 12 attributes(f0-f11), a treatment indicator and 2 binary labels (visits and conversions). After visualizing the correlation values between all the features by a heat map, we find that the treatment is independent of other features in this dataset, fulfilling the independence prerequisite for robust causal inference analysis.

2.3 Meta Learning Methods

Meta-learners are commonly used to address uplift modeling by employing supervised algorithms to predict the effects of treatments and controls. The S-learner is one such approach, utilizing a single model trained on the entire population with a treatment indicator. The uplift is calculated as the difference in predicted outcomes with and without treatment Lo (2002), given by:

$$\tau_i = \mu(X_i, W_i = 1) - \mu(X_i, W_i = 0) \tag{7}$$

where $\mu(X_i, W_i)$ is the model learned by the machine learning algorithm, and we output the probability instead of the binary label to capture the increased conversion rate. Figure 1 illustrates the working mechanism of the S-learner Maksim Shevchenko (2020).

However, the S-learner faces challenges when dealing with high-dimensional covariates, as the model may effectively ignore the treatment effect. For instance, in neural networks, the treatment indicator may receive a near-zero weight, resulting in similar predicted outcomes for both treated and control groups, thereby yielding a zero uplift effect. To mitigate this issue, the T-learner is proposed. The T-learner employs two separate probabilistic models to predict outcomes for the treatment and control groups, respectively. The uplift is determined by subtracting the control group predictions from the treatment group predictions Betlei et al. (2018):

$$\tau_i = \mu_1(X_i, W_i = 1) - \mu_2(X_i, W_i = 0) \tag{8}$$

where μ_1, μ_2 are two models we trained for treatment and control separately. Figure 2 demonstrates the T-learner's operational process Maksim Shevchenko (2020).

The training process:

$$fit\Big(\underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}}_{X_{train}} \underbrace{\begin{matrix} w_1 \\ \cdots \\ w_n \end{matrix}}_{W_{train}}, \underbrace{\begin{matrix} y_1 \\ \cdots \\ y_n \end{matrix}}_{Y_{train}}\Big)$$

The process of applying the model:

$$\begin{matrix} predict \\ proba \end{matrix} \Big(\underbrace{\begin{matrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} \end{matrix}}_{X_{test}} \underbrace{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}_{W_1}\Big) - \begin{matrix} predict \\ proba \end{matrix} \Big(\underbrace{\begin{matrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} \end{matrix}}_{X_{test}} \underbrace{\begin{matrix} 0 \\ \vdots \\ 0 \end{matrix}}_{W_0}\Big) = \underbrace{\begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}}_{uplift}$$

Figure 1: S-learner Uplift Modeling

The training process:

$$model^T = fit\Big(\underbrace{\begin{matrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pk} \end{matrix}}_{X_{train\_treat}}, \underbrace{\begin{matrix} y_1 \\ \cdots \\ y_p \end{matrix}}_{Y_{train\_treat}}\Big), \; model^C = fit\Big(\underbrace{\begin{matrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{q1} & \cdots & x_{qk} \end{matrix}}_{X_{train\_control}}, \underbrace{\begin{matrix} y_1 \\ \cdots \\ y_q \end{matrix}}_{Y_{train\_control}}\Big)$$

The process of applying the model:

$$\begin{matrix} model^T \\ predict \\ proba \end{matrix} \underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} \end{pmatrix}}_{X_{test}} - \begin{matrix} model^c \\ predict \\ proba \end{matrix} \underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} \end{pmatrix}}_{X_{test}} = \underbrace{\begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}}_{uplift}$$

Figure 2: T-learner Uplift Modeling

2.3.1 Training Process

Based on the simplicity and well-performance of the T-learner over other uplift modeling tasks, we decided to use it for our project. The data are split into training and test sets using a 4:6 ratio. Before model training, the data are standardized to have a mean of 0 and a standard deviation of 1. To estimate the potential outcomes, we explore multiple machine learning algorithms, including random forest, decision tree, XGBoost classifier, neural network, LightGBM, and logistic regression. Both grid search and manual parameter adjustments are utilized to optimize the parameters for each algorithm.

2.4 Direct Modeling Method

This section delineates the uplift modeling approach framework employed in our study, which is based on an existing model from the paper Wang et al. (2023), as our model shown in Figure 3. Further, we discuss and apply potential adaptations to enhance the performance of this model. This approach leverages Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) as the foundational architectures, augmented by enhanced feature vectors and causal graphs.

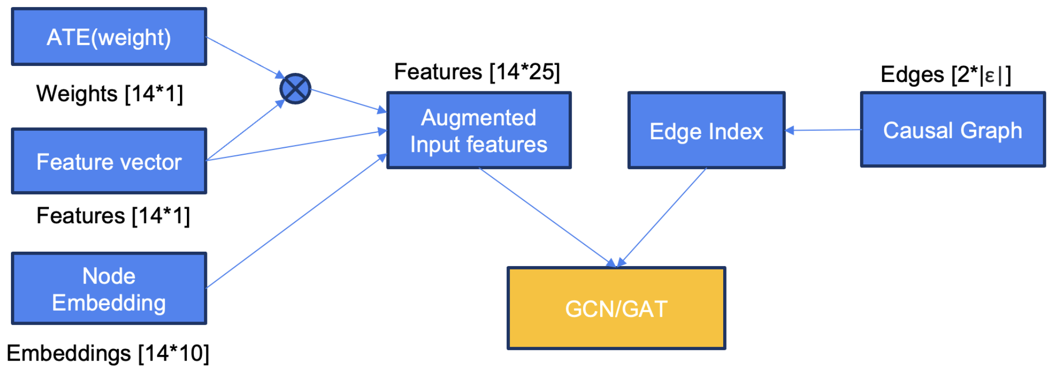


Figure 3: Uplift Modeling

2.4.1 Causal Graph Generation

Causal graphs are integral to the performance of our model. However, the selection of the method for causal graph generation can significantly influence the outcome. To assess the robustness of our model, we evaluated multiple causal graph discovery methods, including Greedy Hill Climbing, as shown in Figure 4 and Greedy Equivalence Search (GES) Alonso-Barba et al. (2011), as shown in Figure 5. Additionally, we introduced perturbations by randomly flipping the edges of the graph to test the model's resilience to structural variations.

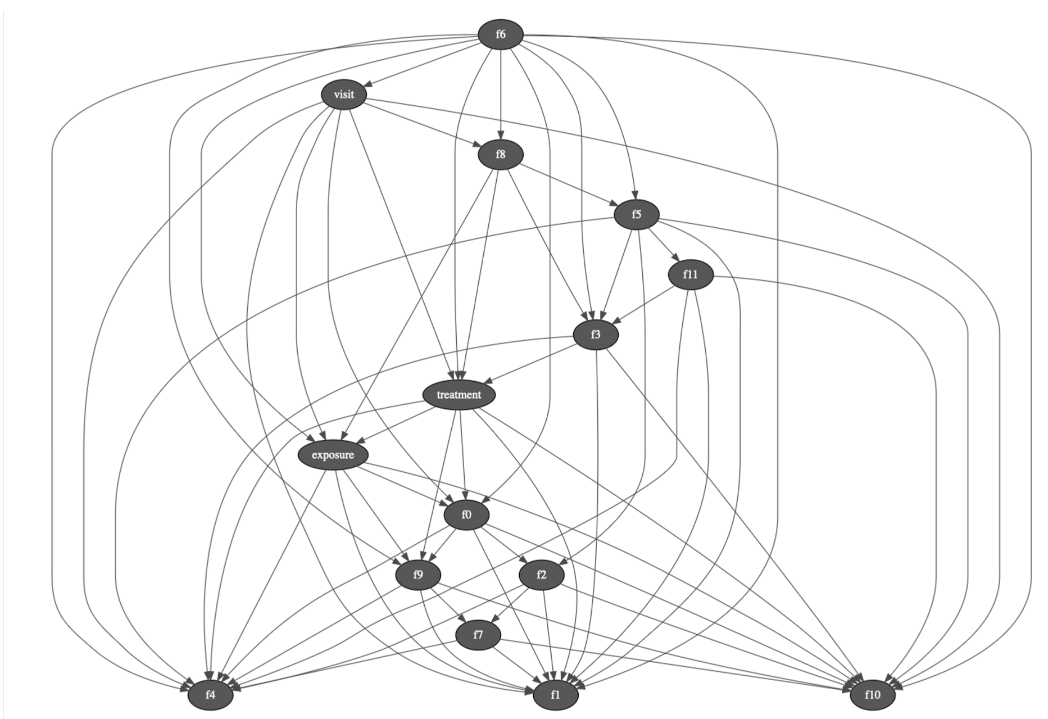


Figure 4: Causal Graph by Greedy Hill Climbing

Randomly Edges Flipping Setting Starting with the causal graph discovered by the Greedy Hill Climbing algorithm, we applied a noise-injection strategy to test the robustness of our model. Each edge in the causal graph (considering both its direction and whether it is one-sided or bidirectional)

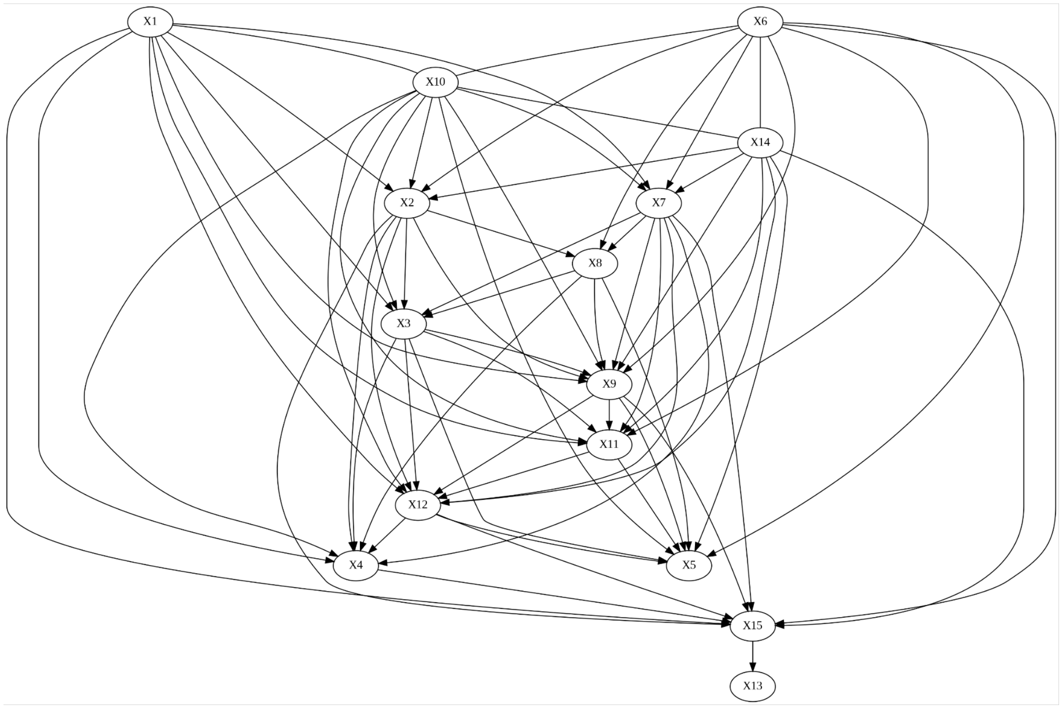


Figure 5: Causal Graph by GES

was subjected to potential modifications. Specifically, 10% of the edges had their causal direction flipped, while an additional 5% of the edges were either added or removed. This procedure was designed to introduce noise into the discovery results, thereby allowing us to evaluate the model's sensitivity to perturbations in the causal graph structure.

2.4.2 ATE-Based Feature Enhancement

To augment the feature vectors, we employed the Average Treatment Effect (ATE) methodology. Specifically, we utilized a Causal Forest Model to compute the ATE for each feature. These ATE values were then used to weight the feature vectors, thereby enhancing them. This enhancement shows great impact on the model performance, as it directly influences the causal weights in the graph neural network models.

2.4.3 Node Embedding

For node embedding, we exploited the sequential properties of the data within our causal graph. Random walks were performed to capture the relational dynamics between nodes, followed by the application of Word2Vec and Node2Vec models to generate the embeddings. Therefore, we capture the correlations among all features by converting the random walk results into node embeddings, thereby enhancing the original input data with its sequential information and representation.

3 Results

3.1 Evaluation Metrics

In uplift modeling, a higher uplift score suggests a greater effect from the treatment (advertising in this context). However, without ground truth, evaluating the model can be challenging. Here we introduce some metrics for model evaluation.

3.1.1 Cumulative Uplift

Rather than looking at the uplift score across all observations, it can be more informative to consider the uplift effect at the top K observations. In practice, decisions about whether to apply treatment are based on the uplift score, making the top K observations critical for determining the cutoff. The formula for cumulative uplift is:

$$\left(\frac{Y^T}{N^T} - \frac{Y^C}{N^C}\right)(N^T + N^C) \tag{9}$$

where Y represents the conversions and N represents the total number of individuals in the group.

3.1.2 Area Under Uplift Curve (AUUC)

The Area Under Uplift Curve (AUUC) is analogous to the area under the curve (AUC) but applied to the uplift curve. Similar to cumulative uplift, the uplift curve is computed for the top k uplift scores, with samples ordered by their uplift score:

$$f(t) = \left(\frac{Y_t^T}{N_t^T} - \frac{Y_t^C}{N_t^C}\right)(N_t^T + N_t^C) \tag{10}$$

where t indicates the top k observations ordered by their uplift score.

3.1.3 Area Under Qini Curve (AUQC)

One limitation of AUUC is its potential bias when the control and treatment groups are unbalanced. The Qini index addresses this by scaling the control group relative to the treatment group:

$$g(t) = Y_t^T - \frac{Y_t^C N_t^T}{N_t^C} \tag{11}$$

Using a similar approach, we compute the Qini curve and the AUQC as follows:

$$f(t) = g(t)\frac{N_t^C + N_t^T}{N_t^T} \tag{12}$$

In this project, we used the normalized AUQC provided by the 'sklift' library Maksim Shevchenko (2020).

3.2 Meta Learning Methods

Table 1 shows the results of the best models we trained using different algorithms. We can see that tree-based classifiers generally perform better than neural networks or simple logistic regression. Notably, the LGBMClassifier achieved the best overall AUUC and AUQC scores. This result is promising, indicating that we can achieve satisfactory outcomes even with a straightforward T-learner, without delving into detailed causal structures. This is useful in real-world scenarios where directors may need quick results from simple models to timely release advertisements and capture market opportunities.

Moreover, the uplift effect observed in the models is relatively small, with an average uplift of around 0.001. This might be due to our focus on conversion rates; while advertising may increase exposure, it does not necessarily lead to immediate actions such as purchases or sign-ups.

3.3 Direct Modeling Method

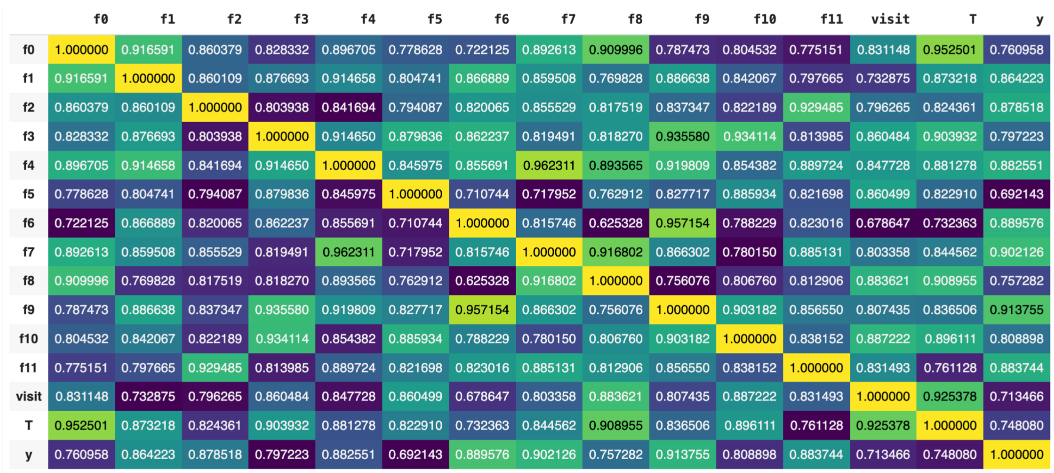
3.3.1 Node Embedding Analysis

We conducted an analysis of the node embeddings generated from the Node2Vec on Random Walk results. The node embeddings were evaluated based on their pairwise similarity, providing insights

Model	AUUC	AUQC	uplift at top 30%	Weighted average uplift
DecisionTree	0.872	0.002	0.001	0.001
XGBClassifier	0.845	0.163	0.004	0.001
MLPClassifier	0.637	0.019	0.002	0.001
RandomForestClassifier	0.750	0.119	0.003	0.001
LGBMClassifier	0.872	0.179	0.004	0.001
LogisticRegression	0.665	0.079	0.003	0.001

Table 1: Uplift Modeling Results

into the relational structure of the graph. The distance map, as shown in Figure 6, illustrates the similarity among node embeddings. Higher similarity scores correspond to stronger associations, which are indicative of the underlying causal relationships in the data.



	f0	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	visit	T	y
f0	1.000000	0.916591	0.860379	0.828332	0.896705	0.778628	0.722125	0.892613	0.909996	0.787473	0.804532	0.775151	0.831148	0.952501	0.760958
f1	0.916591	1.000000	0.860109	0.876693	0.914658	0.804741	0.866889	0.859508	0.769828	0.886638	0.842067	0.797665	0.732875	0.873218	0.864223
f2	0.860379	0.860109	1.000000	0.803938	0.841694	0.794087	0.820065	0.855529	0.817519	0.837347	0.822189	0.929485	0.796265	0.824361	0.878518
f3	0.828332	0.876693	0.803938	1.000000	0.914650	0.879836	0.862237	0.819491	0.818270	0.935580	0.934114	0.813985	0.860484	0.903932	0.797223
f4	0.896705	0.914658	0.841694	0.914650	1.000000	0.845975	0.855691	0.962311	0.893565	0.919809	0.854382	0.889724	0.847728	0.881278	0.882551
f5	0.778628	0.804741	0.794087	0.879836	0.845975	1.000000	0.710744	0.717952	0.762912	0.827717	0.885934	0.821698	0.860499	0.822910	0.692143
f6	0.722125	0.866889	0.820065	0.862237	0.855691	0.710744	1.000000	0.815746	0.625328	0.957154	0.788229	0.823016	0.678647	0.732363	0.889576
f7	0.892613	0.859508	0.855529	0.819491	0.962311	0.717952	0.815746	1.000000	0.916802	0.866302	0.780150	0.885131	0.803358	0.844562	0.902126
f8	0.909996	0.769828	0.817519	0.818270	0.893565	0.762912	0.625328	0.916802	1.000000	0.756076	0.806760	0.812906	0.883621	0.908955	0.757282
f9	0.787473	0.886638	0.837347	0.935580	0.919809	0.827717	0.957154	0.866302	0.756076	1.000000	0.903182	0.856550	0.807435	0.836506	0.913755
f10	0.804532	0.842067	0.822189	0.934114	0.854382	0.885934	0.788229	0.780150	0.806760	0.903182	1.000000	0.838152	0.887222	0.896111	0.808898
f11	0.775151	0.797665	0.929485	0.813985	0.889724	0.821698	0.823016	0.885131	0.812906	0.856550	0.838152	1.000000	0.831493	0.761128	0.883744
visit	0.831148	0.732875	0.796265	0.860484	0.847728	0.860499	0.678647	0.803358	0.883621	0.807435	0.887222	0.831493	1.000000	0.925378	0.713466
T	0.952501	0.873218	0.824361	0.903932	0.881278	0.822910	0.732363	0.844562	0.908955	0.836506	0.896111	0.761128	0.925378	1.000000	0.748080
y	0.760958	0.864223	0.878518	0.797223	0.882551	0.692143	0.889576	0.902126	0.757282	0.913755	0.808898	0.883744	0.713466	0.748080	1.000000

Figure 6: Node Embedding Distance Map

The distance map analysis reveals that our method effectively captures the relational dynamics within the data. The embeddings generated through Node2Vec exhibit coherent patterns, reflecting the true causal structure more accurately compared to traditional methods.

3.3.2 Performance Analysis

We evaluated the performance of our uplift modeling approach using the Area Under the Uplift Curve (AUUC) metric. Various feature processing methods were compared, including the use of original data features and those augmented with causal weighting. The performance metrics for different models are summarized in Table 2.

Method	GCN	GAT
Hill Climbing (Feature only)	0.721959	0.794647
Hill Climbing (Causal Weighting)	0.733972	**0.836097**
GES (Causal Weighting)	**0.743254**	0.824441
Random Flipping (Causal Weighting)	0.549209	0.657535
Random	0.499138	0.499138

Table 2: Uplift Modeling Results (AUUC)

The results indicate that utilizing GES for causal graph discovery significantly enhances performance compared to only using augmented input data (Causal Weighting). Introducing noise through random edge flipping considerably degrades performance, underscoring the critical importance of accurate causal graph discovery. These findings suggest that simplistic methods like hill climbing are insufficient and yield unstable results, highlighting the necessity for robust graph generation techniques.

4 Discussion and Limitations

This study addresses the challenges and limitations encountered in uplift modeling, particularly concerning the unavailability of ground truth labels, the risk of model overfitting, and the nuances of meta-learning and direct modeling methods.

4.1 Ground Truth in Uplift Modeling

A fundamental challenge in uplift modeling is the absence of direct ground truth labels, as an individual cannot simultaneously be in both the treatment and control groups. This unique aspect necessitates the use of causal inference to simulate unobserved outcomes, transforming it into a scenario akin to a typical machine learning problem where features predict targets. However, without true ground truth, validation becomes complex, as it involves comparing model predictions against estimated rather than actual outcomes. Synthetic data can aid this process by simulating both scenarios, providing a clearer but still imperfect proxy for real-world data. This inherent limitation underscores the difficulty in achieving accurate validation and highlights the need for robust simulation techniques to approximate real-world conditions.

4.2 Model Overfitting

In striving to capture intricate data patterns within the training dataset, there is a significant risk of overfitting, where the model becomes overly specialized to the training data, including its noise and outliers. This results in diminished performance on unseen, new data as the model loses its generalizability. This is particularly salient in causal inference models, where the drive to meticulously model data interactions can inadvertently lead to models that are too complex and less effective at predicting new instances. Techniques like cross-validation, regularization, and model simplification are critical strategies employed to mitigate overfitting, enhancing the model's ability to perform reliably across diverse datasets.

4.3 Meta-Learning Method Challenges

While meta-learning methods offer a powerful tool for uplift modeling, they come with their own set of limitations. The effectiveness of these methods depends heavily on the chosen model's ability to capture the true patterns and relationships within the data. Limited experimentation and an insufficient exploration of potential models might prevent the identification of an optimal model. Moreover, meta-learning algorithms can sometimes fail to accommodate the complex structural dynamics of the data, either due to an overly simplistic approach or constraints within the algorithm's design and parameter settings. These limitations can restrict the performance and applicability of meta-learning methods in accurately modeling uplift.

4.4 Direct Modeling Method Sensitivities

In the direct modeling approach, the setting of initial causal weights is critically sensitive, as inappropriate initial values can significantly affect the model's convergence and accuracy. If these weights are not close to their optimal settings, the model may converge slowly or become trapped in local minima, failing to capture the true causal relationships effectively. Additionally, incorrect initial settings of these weights can result in biased estimates and reduced statistical efficiency, undermining the credibility and utility of the model.

These limitations highlight the need for continuous refinement and testing of uplift modeling techniques to ensure they are both robust and adaptable to varying data environments and that they maintain a balance between complexity and practicality.

5 Conclusion and Future Work

5.1 Conclusion

This study explored the application of uplift modeling techniques within the realm of online advertising, focusing particularly on the estimation of the advertising lift effect. We employed both meta-learning and direct modeling methods to estimate the individual treatment effects on the Criteo Uplift Prediction Dataset. The comparative analysis was conducted using the AUUC and AUQC metrics to gauge the efficacy of these methods.

Our results indicate that meta-learning methods, particularly the T-learner, performed robustly, offering a straightforward yet effective means of predicting the uplift effect. This approach proved particularly beneficial for decision-makers needing rapid results to inform timely marketing strategies. The T-learner, facilitated by tree-based classifiers such as the LGBMClassifier, achieved notable performance, underscoring the utility of meta-learning in scenarios where causal relationships are less explicit or complex.

Conversely, the direct modeling method integrating Graph Neural Networks with causal knowledge demonstrated the potential for refining uplift predictions by capturing complex data interactions more effectively. The causal graph techniques, including the novel node embedding processes and ATE-based feature enhancements, significantly improved the precision of uplift estimates. These methods were particularly adept at handling the inherent complexities of online user behavior and the impact of advertising, which traditional models might overlook.

However, both approaches revealed inherent limitations such as sensitivity to model configurations and data quality. For instance, the robustness test involving random edge flipping in causal graphs illustrated a decrease in model performance, suggesting that our model's effectiveness is contingent upon the accurate representation of causal relationships.

In summary, our study underscores the importance of choosing appropriate uplift modeling techniques based on the specific requirements and constraints of marketing campaigns. By continuously refining these models and integrating new advancements in machine learning and causal inference, businesses can significantly enhance their strategic decision-making processes, leading to more targeted and effective advertising strategies.

5.2 Future Work

Looking ahead, our research will extend into several promising avenues within uplift modeling to enhance both methodological robustness and practical applicability across various digital marketing contexts.

Firstly, we aim to broaden our exploration of meta-learning techniques beyond the T-learner implemented in this project. Future studies will include the S-learner, known for its simplicity and the integration of the treatment indicator as a feature, which simplifies model architecture while facilitating easier management and maintenance. Additionally, the R-learner will be evaluated for its ability to minimize bias through a focus on residual analysis, enhancing stability and reliability in treatment effect estimations. We also plan to incorporate the X-learner, which is particularly adept at managing imbalanced datasets by cross-estimating treatment effects, thus improving accuracy and reducing variance. By integrating these methods—S-learner, R-learner, and X-learner—with our current T-learner approach, we anticipate developing a more comprehensive framework that leverages the unique strengths of each method to provide more precise and dependable causal effect estimates.

In the realm of direct modeling, we propose to refine our use of Graph Neural Networks (GNNs) by integrating edge weights from causal discovery into the causal graphs used as inputs for GCN/GAT models. This modification aims to transform the current binary representation into a more nuanced numerical format, reducing the impact of potential inaccuracies in the causal graph construction. By employing these edge weights, we can enhance the granularity and accuracy of the relationships represented in the graph, enabling the GNNs to more effectively learn and generalize from complex datasets. This approach is expected to bolster the robustness and precision of our models, particularly in dynamically changing environments typical of online advertising spaces.

Together, these developments will advance our understanding of uplift modeling's capabilities and limitations, contributing to more targeted, effective, and economically efficient advertising strategies in the digital era. By continuously integrating advanced machine learning techniques and exploring their applications in real-world scenarios, our research will strive to bridge the gap between theoretical models and practical marketing needs.

5.3 Job Assignment

In this project, Xiaotian Shao and Jingyi Zhou are responsible for meta learning method while Jintong Luo and Can Chen are responsible for direct methods.

References

Gediminas Adomavicius and Alexander Tuzhilin. 1997. Discovery of actionable patterns in databases: The action hierarchy approach. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 111–114.

J. I. Alonso-Barba, L. D. L. Ossa, J. A. Gamez, and J. M. Puerta. 2011. Scaling up the greedy equivalence search algorithm by constraining the search space of equivalence classes. *International Journal of Approximate Reasoning*, pages 194–205.

Artem Betlei, Eustache Diemert, and Massih-Reza Amini. 2018. Uplift prediction with dependent feature representation in imbalanced treatment and control conditions. In *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part V 25*, pages 47–57. Springer.

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24:123–140.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Tianxi Cai, Tian Lu, Peggy H. Wong, et al. 2011. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282.

Floris Devriendt, Darie Moldovan, and Wouter Verbeke. 2018. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data*, 6(1):13.

Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. 2018. A large scale benchmark for uplift modeling. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2):193–225.

Behram Hansotia and Brad Rukstales. 2002. Incremental value modeling. *Journal of Interactive Marketing*, 16(3):35–46.

Jennifer L. Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational & Graphical Statistics*, 20(1):217–240.

Shin Kawanaka and Daisuke Moriwaki. 2019. Uplift modeling for location-based online advertising. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Recommendations, Geosocial Networks and Geoadvertising*, pages 1–4.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2009. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18:140–181.

Randall A. Lewis, Justin M. Rao, and David H. Reiley. 2011. Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th International Conference on World Wide Web*.

CH Liu and Benjamin Paul Chamberlain. 2018. Online controlled experiments for personalised e-commerce strategies: Design, challenges, and pitfalls. arXiv preprint arXiv:1803.06258.

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

Irina Elisova Maksim Shevchenko. 2020. User guide for uplift modeling and casual inference. https://www.uplift-modeling.com/en/latest/user_guide/index.html.

Nicholas Radcliffe and Patrick Surry. 1999. Differential response analysis: Modeling true responses by isolating the effect of a single action. In *Credit Scoring and Credit Control IV*.

Nicholas J. Radcliffe and Patrick D. Surry. 2011. Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions.

Vanguarde. 2022. Study Series: Uplift Modeling. https://www.kaggle.com/code/vanguarde/study-series-uplift-modeling. Accessed: [insert date of access].

Hal R. Varian. 2016. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7310–7315.

Haowen Wang, Xinyan Ye, Yangze Zhou, Zhiyi Zhang, L. Zhang, and Jing Jiang. 2023. Uplift modeling based on graph neural network combined with causal knowledge. *ArXiv*.