Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230930

Multi-treatment casual analysis using improved meta learner and uplift tree

Keyu Hong[1], Wen Chen[2], Ruidong Luo[3] and Andy Zhu[4]

[1]Department of Mathematics & Statistics, McMaster University, Hamilton, L8S 4S4, Canada, hongkeyu95@gmail.com
[2]The Department of Philosophy, Xiamen University, Xiamen, 361005, China, 18392679681@163.com
[3]Department of Material Science and Technology, Beijing University of Chemical Technology, Beijing, 100029, China, 17801198583 @163.com
[4]NA, St. George's School, Vancouver, V6S 1V1, Canada, andyzyzhu23@gmail.com

Abstract. Causality is an appealing but challenging domain for researchers in generations. Recently, researchers have shifted their focus to combining traditional causal inference methods and machine learning models to get both advantages. Meta learner is an algorithm for causal inference, including T-learner, S-learner, and X-learner. Another popular way in causal inference is based on decision tree learning, one of the predictive modeling approaches. Many existing works focus on estimating the causal effect of binary treatment. However, there are also many cases in the real world when the treatment has more than two values. These methods cannot be used directly in multivalued treatment cases. According to the mathematization of causality, we improved the binary meta-learner process to be applicable in multi-treatment situations. At the same time, we also preliminarily explored the technique of uplifting trees. Finally, we applied the two methods to analyze parents' and children's learning situations in hundreds of families to test the effect of improvement.

Keywords: muti-treatment, meta learner, uplift tree.

1. Introduction

There is a well-known statement: correlation (or, more generally, statistical association) does not imply causation. In statistics, correlation is a relationship between two variables representing an increasing or decreasing trend [1]. Causation indicates that the cause is partly responsible for the effect, and the result is also somewhat dependent on the cause [2].

The crucial part of a causation study is to reduce the bias within. Randomized control trials are a compelling solution for estimating causal effects because this makes the control and treatment groups comparable. However, randomized experiments cannot be used in every situation as they could be time-consuming, expensive, and sometimes infeasible [3].

When studying causal inference using observational data, traditional methods, including matching, propensity score, subclassification, weighting, and doubly robust estimation, have been proposed to acquire covariate balance across treatment groups [3,4,5,6]. With the rapid development of computer science, traditional machine learning methods, like meta-learner and uplift tree, have been used to estimate causal effects.

DOI: 10.54254/2755-2721/6/20230930

Generally, the meta-learning-based algorithms have 2 steps: estimate the conditional mean outcome, and the prediction model learned in this step is the base learner, then derive the CATE estimator based on the difference of results obtained from the first step. Existing meta-learning methods include S-learner, T-learner, X-learner, U-learner, and R-learner [7].

One of the predictive modeling methods based on a decision tree is another popular approach in causality studying. The decision tree is a nonparametric supervised learning method for classification and regression. The purpose is to develop a model that predicts the value of a target variable by learning simple decision rules inferred from data. The tree-based framework also can be extended to uni- or multi-dimensional treatments [8]. Each dimension can be discrete or continuous.

In this paper, we utilize the traditional machine learning method, meta-learner, and tree-based algorithms to explore the causal effects we are interested in within a dataset about student alcohol consumption. Besides, we also improved the meta-learner to study the causal impact of multi-valued treatment, and we found that tree-based methods perform well when the outcome is binary. The meta-learning plans fit the situation when the product is continuous.

For this research, CausalML was used. CausalML is a python package that allows users to estimate Conditional Average Treatment Effect (CATE) or Individual Treatment Effect (ITE) from a given data. Using machine learning algorithms, this package contains uplifting modeling and causal inference methods. However, its implementation of multiple treatments is flawed as it compares treatment with the control group pairwise, neglecting other treatment groups, which is statistically problematic. Our implementation of X learner would try and fix that.

2. Preliminaries

We use the Neyman-Rubin potential outcome framework and assume a distribution $\mathcal{P}$. Thus, we have $(\mathrm{Y_i}(0),\ \mathrm{Y_i(k)},\ \mathrm{X_i}, \mathrm{T_i}) \sim \mathcal{P}$, where $\mathrm{X_i} \in \mathbb{R}^{\mathrm{d}}$ is a d-dimensional real-value feature vector, $\mathrm{T_i} \in \{0, 1, \ldots, \mathrm{k}\}$ is the treatment-assignment indicator, $\mathrm{Y_i}(0) \in \mathbb{R}$ is the potential outcome of unit i, when i is assigned to the control group, then $\mathrm{Y_i(k)} \in \mathbb{R}$ is the likely outcome of unit i, when i is given to the treatment group k. With this definition, the ATE between treatment groups m and n is defined as

$$ATE \coloneqq \mathbb{E}[Y(m) - Y(n)].$$

Furthermore, we have the following representation of $\mathcal{P}$:

$$X \sim \Lambda,$$

$$Y(k) = \mu_k(X) + \varepsilon(k),$$

Where Λ is the marginal distribution of X, $\varepsilon(\mathrm{k})$ is the zero-mean random variables independent of X and W. The ITE of unit i, $\mathrm{D_i}$, for treatment groups m and n is defined as

$$D_i \coloneqq Y_i(m) - Y_i(n).$$

The CATE of unit i, between treatment groups m and n would then be

$$\tau(x) \coloneqq \mathbb{E}[D \mid X = x] = \mathbb{E}[Y(m) - Y(n) \mid X = x]$$

2.1. Introduction to meta-learners

We will give a brief introduction to some proposed meta-learners. For this part, we would only consider binary treatment situations.

T-learner can be broken down into two steps. First, we use one model per treatment variable. The response function would then be,

$$\mu_0 = \mathbb{E}[\mathrm{Y}|\mathrm{T} = 0, \mathrm{X}],$$

$$\mu_1 = \mathbb{E}[\mathrm{Y}|\mathrm{T} = 1, \mathrm{X}],$$

where i is the complementary treatment. We could use any machine-learning model to train on observations in each treatment group. We denote the estimated function as $\hat{\mu}_1$ and $\hat{\mu}_0$. Then, we obtain

DOI: 10.54254/2755-2721/6/20230930

2 different models in total. Second, the CATE desired between the controlled group, and the treated group is then obtained as

$$\hat{\tau}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x).$$

S-learner is the most intuitive and the simplest learner we have. We use a single machine-learning model, $\hat{\mu}$, to estimate the response function,

$$\mu = \mathbb{E}[Y|T, X].$$

For S-learner, we treat this as a regular machine-learning problem and include our treatment, T, as a feature in the model. Then, we can predict outcomes under different treatments and obtain our CATE between treatment i and treatment j as

$$\widehat{\tau_{ij}}(x) = \hat{\mu}(X, T = j) - \hat{\mu}(X, T = i).$$

Note that S-learner would work on both binary treatment problems and discrete and continuous treatments. However, S-learner tends to bias the CATE toward zero, which will be shown later in this paper with real-world data.

X-learner is a significantly more complex algorithm than the previous two and would also be the primary focus of this paper. X-learner can be broken down into two stages. First, estimate the response function,

$$\mu_0 = \mathbb{E}[Y|T = 0, X],$$

$$\mu_1 = \mathbb{E}[Y|T = 1, X],$$

Where μ_0 is the response function for the controlled group and μ_1 is the response function for the treated group. We could estimate this response function using any machine learning models. We denote the estimated functions as $\hat{\mu}_0$ and $\hat{\mu}_1$. Second, we calculate the treatment effects for each treatment group using the estimators obtained above.

$$\widehat{\tau_0}(x) = \hat{\mu}_1(X, T = 0) - Y(0)$$

$$\widehat{\tau_1}(x) = Y(1) - \hat{\mu}_0(X, T = 1)$$

Then, with the estimated CATE, we can directly estimate the CATE with complete input features. The effect functions are then,

$$\hat{\mu}_{\tau_0} = \mathbb{E}[\hat{\tau}(X)|T = 0]$$

$$\hat{\mu}_{\tau_1} = \mathbb{E}[\hat{\tau}(X)|T = 1]$$

Intuitively, X-learner can be broken down into two steps. In the first step, we fill in the unobserved data with predictions of counterfactuals like the T-learner, and in the second stage, we use our completed data to estimate CATE.

In the following sections, propensity scores will also be utilized. Rosenbaum and Rubin first introduced it as the probability of treatment assignment conditional on the characteristics of the sample data [9]. Mathematically speaking, it is defined as $P(T|X)$, where X is the characteristics of the sample and T is the treatment the model received. The propensity score would be used to balance the covariate distribution across different treatment groups, which could lead to an inaccurate weight of other treatment groups. This topic will be further discussed in the next section.

2.2. Tree-based algorithm

2.2.1. Uplift tree

For the uplift model, predictive features are required for $\mathrm{HTE}(Y_t - Y_c)$, and feature selection methods based on the uplift tree tend to choose more important features for HTE.

DOI: 10.54254/2755-2721/6/20230930

Let us formalize the problem.

Like building the ordinary decision tree, we hope that we choose the proper splitting criteria, which can maximize the information gained before and after the split. However, there is a little bit of difference that we hope to maximize the distribution gap of outcomes between the treatment group and control group.

That means if we define

$$D(P:Q)$$

to measure the distribution divergence between P and Q, our purpose is to maximize this formula:

$$D_{gain}(A) = D\left(P^T(Y):P^C(Y)\middle|A\right) - D(P^T(Y):P^C(Y))$$

The definition of conditional divergence $D\left(P^T(Y):P^C(Y)\middle|A\right)$:

Let A the splitting condition and N be the number of samples. After splitting, let a be one of the outcomes and $N(a)$ be the number of the remaining pieces.

We have:

$$D\left(P^T(Y):P^C(Y)\middle|A\right) = \sum_a \frac{N(a)}{N} D(P^T(Y|a):P^C(Y|a))$$

When the tree has been built, we calculate the treatment effect.

We know that the current leaf nodes are subgroups of objects for which the treatment class distribution differs from the control class distribution.

Now we analyze the case of binary outcomes.

Suppose Y will just be 0 or 1, the later outcome is what we want.

For a new sample it meets the condition to leaf node l, the treatment effect will be:

$$P^T(Y=1|l) - P^C(Y=1|l)$$

If y is a continuous variable, the treatment effect will be:

$$E^T(Y=1|l) - E^C(Y=1|l)^{[11]}$$

2.2.2. Causal forest

This algorithm uses the ensemble method on some built uplift trees and then averages the treatment effect of each canal tree (uplift tree).

In addition, the algorithm needs to meet a hypothesis, which is

$$T \perp Y|X$$

It means that treatment and outcome will be independent after controlling all confounders X [12].

3. Proposed methods of meta learner

Realistic needs often go beyond the scope of binary treatment. This section will introduce our improved models of S Learner, T Learner, and X Learner. It is worth noting that when looking into the causal relationship between variables in the multiple-treatment case, it is statistically incorrect to simply give a pairwise comparison like the binary case, for it fails to utilize the data of other treatment groups.

3.1. S and T learner

Extending the binary version of S and T learners to multiple-treatment models is relatively trivial. Since the S learner directly uses the treatment as a feature in the binary case, the same principle can be applied to the multiple-treatment case.

For T learners, similar to the binary case, a model would be trained for each group based on factual data. These models would be used to predict counterfactual results and calculate treatment effects as described in the following formula:

DOI: 10.54254/2755-2721/6/20230930

$$\tau(X_i) = M_1(X_i) - M_2(X_i)$$

Where τ is the treatment effect, M_1 and M_2 are the models fitted based on treatment groups 1 and 2, and X_i is the characteristic of a specific set of sample data.

3.2. X learner

Improvements of the X Learner are more complicated. As explained earlier, the individual models trained for a T-Learner would be used as an intermediate result to predict counterfactual treatment effects. Here an example is given on comparing treatment groups 0 and 1 in a 3-treatment experiment:

$$\tau(X, T = 0) = M_1(X, T = 0) - Y_{T=0}$$

$$\tau(X, T = 1) = Y_{T=1} - M_0(X, T = 1)$$

$$\tau(X, T = 2) = M_1(X, T = 2) - M_0(X, T = 2)$$

Here T is the treatment effect that can take 0, 1, 2. M_1 and M_0 are models fit in the T learner based on data of treatment groups 0 and 1, respectively. Y denoted the factual data from a specific treatment group specified by its subscript.

Similar to the binary case, we fit three models on $\tau(X, T = 0), \tau(X, T = 1), \text{and } \tau(X, T = 2)$ to predict the treatment effect on a specific treatment group.

We then fit a model on X to predict the propensity score e. The advantages of doing so have been discussed earlier. Using e as a weight for τ of each treatment group, we can expect the final treatment effect:

$$\tau(X) = e_0 M_{\tau,T=0}(X) + e_1 M_{\tau,T=1}(X) + e_2 M_{\tau,T=2}(X)$$

In this way, the treatment effect of groups 0 and 1 could be estimated while accurately considering the cases in which T=2. This example could be extended to more than three treatment groups following the same manner when dealing with treatment group 2: using the trained model to predict the treatment effect counterfactually and using their propensity scores as the weight in the final formula.

4. Real-world experiments

4.1. Application of meta learner

In this section, we will perform our analysis on a real-life dataset looking into the relationship between the alcohol consumption of students and their grades [10]. The following graphs are some exploratory data analyses of our dataset.

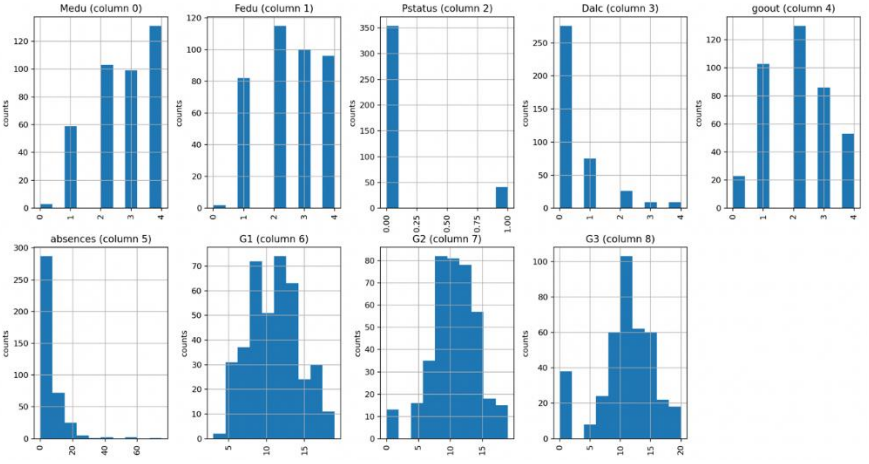

Figure 1. Column distribution.

DOI: 10.54254/2755-2721/6/20230930

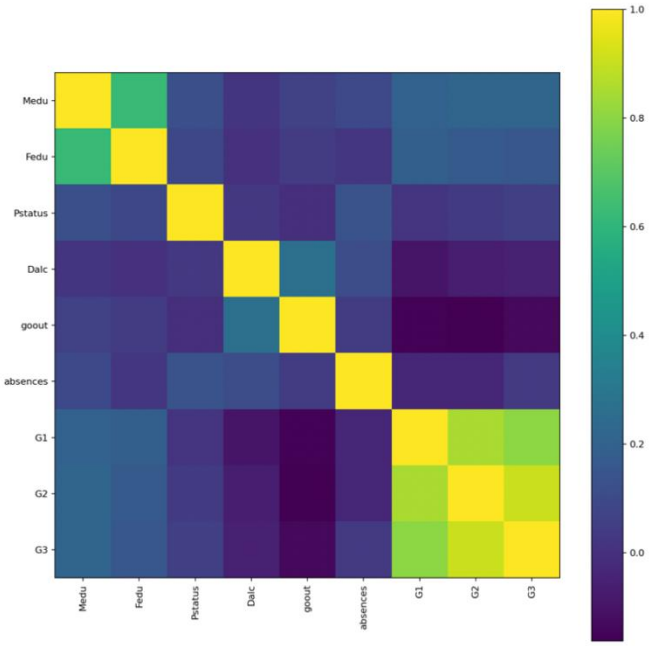

Figure 2. Correlation matrix.

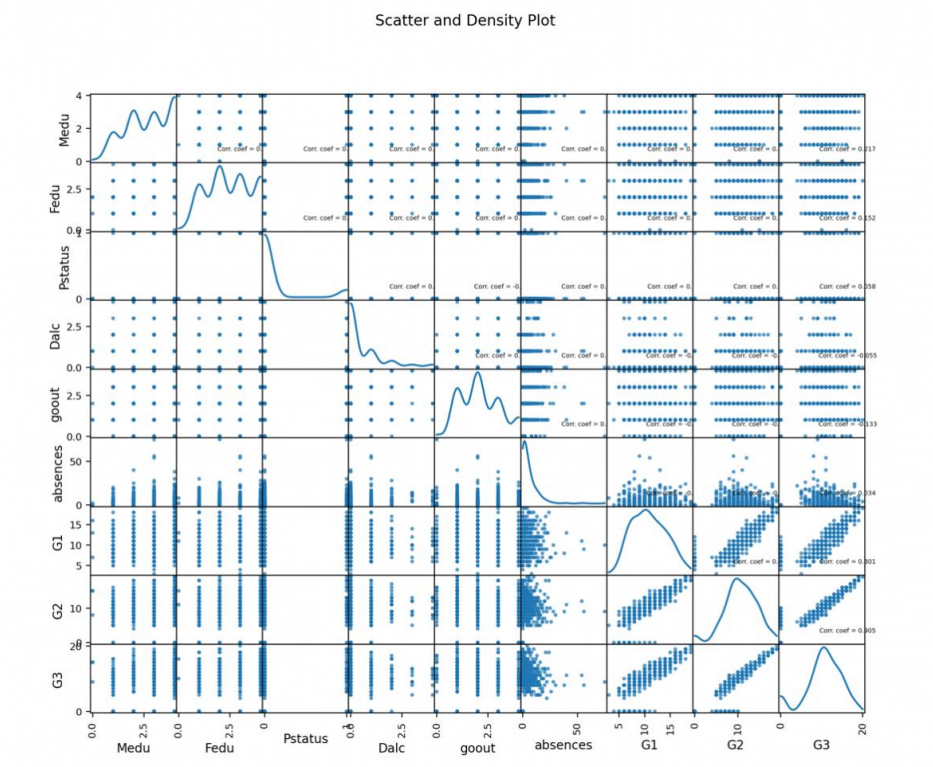

Figure 3. Scatter matrix.

DOI: 10.54254/2755-2721/6/20230930

The Average Treatment Effect (ATE) between many data pairs is calculated using the models proposed in the previous section. It is worth mentioning that there aren't any fixed specific "correct" solutions to the answers. We are using this dataset to show that our model could produce results compatible with an existing package.

We would apply traditional machine learning methods, X Learner, implemented by a package and our implementation on this dataset.

Using the improved X learner, we looked into the causal relationship between the level of mother's educational level with the number of absences in math class, the level of the father's educational level with the number of absences in math class, the workday alcohol consumption with the number of absent days and the final grade of their math classes. The result is shown in the following graph. The education levels are evaluated based on 0 to 4: 0 being the low level and 4 being the high level. At the same time, alcohol consumption is estimated from 1-5, 1 being never drinking and five being consuming a considerable amount of alcohol. Further details can be found on the website.

Table 1. ATE of parents' education level and number of absent days using our model.

ATE	Group 1-0	Group 2-0	Group 3-0	Group 4-0
Mother's education level-absences	2.9514942	4.865572	7.233008	5.1655226
Father's education level -absences	7.4777775	2.8022387	4.725139	3.1205428

Table 2. ATE of parents' education level and number of absent days using package.

ATE	Group 1-0	Group 2-0	Group 3-0	Group 4-0
Mother's education level-absences	3.1269537	5.83861463	5.86365847	5.68739891
Father's education level -absences	5.42818505	1.49965468	2.67241615	1.59391202

Table 3. ATE of alcohol consumption on workday with absences and final grade using package.

ATE	Group 2-1	Group 3-1	Group 4-1	Group 5-1
Alcohol Consumption - absences	1.88663458	0.61650378	4.00502642	-0.390577
Alcohol Consumption -final grade (G3)	-9.23595804	2.5338911	1.1233018	3.480226

Table 4. ATE of alcohol consumption on workday with absences and final grade using our model.

ATE	Group 2-1	Group 3-1	Group 4-1	Group 5-1
Alcohol Consumption - absences	1.0758184	1.8308517	2.9469342	1.050765
Alcohol Consumption -final grade (G3)	-0.26763344	0.59105015	0.0579866	0.4181542

We could observe that a higher education level compared with level 0 (none) might lead to a higher number of absent days, which is against our intuition and might be due to the lack of casual relationships or the insufficiency of the data. So is the case for relationships between alcohol consumption on a workday and final grades or absences.

We could also observe that the result of our model is slightly different from the package implemented result, even using the same base regressor (XGBRegressor) and the same hyperparameters. This might

DOI: 10.54254/2755-2721/6/20230930

result from the differences in propensity score selection or the deficiencies of the package as it forces the user to choose a control group and compares the control group and multiple treatment groups. The latter is inconvenient as it fails to provide users with casual relationships between different treatment groups and is problematic as it performs pairwise comparisons similar to the binary case neglecting other treatments, which is statistically inaccurate.

4.2. Tree-based algorithm

Consider treatment is mother job (Mjob), and the outcome is whether the teenager wants to take higher education(higher).

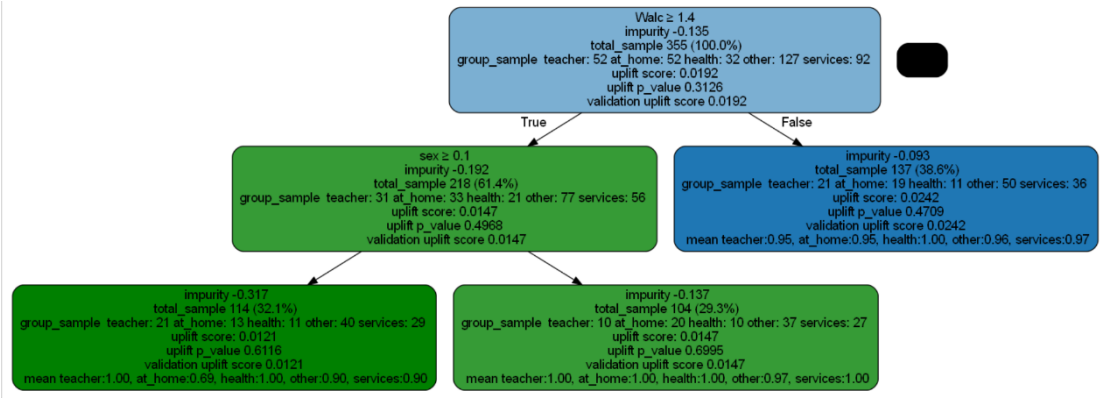

Figure 4. We can see that blue users (Walc<1.4) react strongly to the treatment, while green users react slightly.

Consider treatment is family educational support (farms up), and outcome is whether the teenager wants to take higher education(higher).

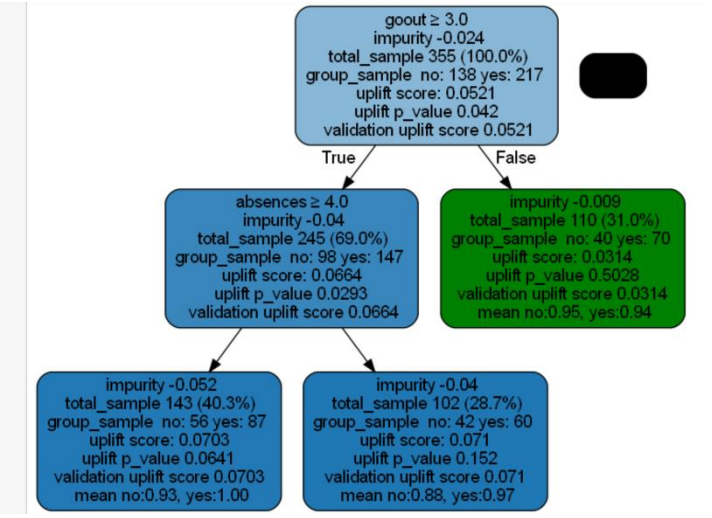

Figure 5. We can see that blue users (gout>=3.0, absences>=4.0) react strongly to the treatment, while green users react slightly to it.

5. Conclusion

In this essay, we implemented our version of X learner that could deal with multiple treatments. While an existing package (CasualML) is statistically flawed as it implements various treatment cases by comparing treatments pairwise, neglecting the weight of the other unstudied treatment groups. Our code managed to fix that mistake. We tested our model on a real-life data set regarding student alcohol

DOI: 10.54254/2755-2721/6/20230930

consumption and yielded different ATE. Though insufficient to claim any casual relationships, the application proved that our version of X Learner managed to avoid the mistake and estimate ATE reflecting all available data. When we consider the impact of treatment on the subjective inclination or the causal effect on whether or not you are willing to do something, the uplift tree can perform very well. However, the specific implementation cannot estimate multiple treatments' causal effects, and there is still much space for improvement.

Acknowledgment
Keyu Hong and Wen Chen contributed equally to this work and should be considered co-first authors.

References

[1] Altman, Naomi, and Martin Krzywinski. "Points of Significance: Association, correlation, and causation." Nature methods 12.10 (2015).
[2] Yao, Liuyi, et al. "A survey on causal inference." ACM Transactions on Knowledge Discovery from Data (TKDD) 15.5 (2021): 1-46.
[3] Alves, M. F. "Causal inference for the brave and true." (2021).
[4] Lopez, Michael J., and Roee Gutman. "Estimation of causal effects with multiple treatments: a review and new ideas." Statistical Science (2017): 432-454.
[5] Stuart, Elizabeth A. "Matching methods for causal inference: A review and a look forward." Statistical science: a review journal of the Institute of Mathematical Statistics 25.1 (2010): 1.
[6] Scotina, Anthony D., and Roee Gutman. "Matching algorithms for causal inference with multiple treatments." Statistics in medicine 38.17 (2019): 3139-3167.
[7] Yao, Liuyi, et al. "A survey on causal inference." ACM Transactions on Knowledge Discovery from Data (TKDD) 15.5 (2021): 1-46.
[8] Wang, Pengyuan, et al. "Robust tree-based causal inference for complex ad effectiveness analysis." Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. 2015.
[9] Rosenbaum, Paul R. and Donald B. Rubin. “The central role of the propensity score in observational studies for causal effects.” Biometrika 70 (1983): 41-55.
[10] Student Alcohol Consumption, Kaggle, https://www.kaggle.com/datasets/uciml/student-alcohol-consumption?select=student-mat.csv.
[11] Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. Knowl. Inf. Syst., 32(2):303–327, August 2012.
[12] Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests[J]. Journal of the American Statal Association, 2018, 113(523):1228-1242.