Using Machine Learning for Causal Inference

Yue Wu
School of Economics and Management
Tsinghua University
yue-wu24@mails.tsinghua.edu.cn

Abstract

Causal inference, a hot topic in modern statistics, aims to estimate the causal effects of treatments. With the growing availability of high-dimensional, large-scale data in industrial scenarios such as targeted advertising, machine learning has become a key tool in causal inference, especially for modeling heterogeneous treatment effects. This paper reviews methodology of applying machine learning methods in causal scenarios. Firstly, we review basic concepts and assumptions in causal inference, as well as some important topics such as covariate balance and propensity score. Then, we systematically introduce a range of ML-based causal estimators, such as meta-learners, causal forests, and double machine learning. Furthermore, we discuss the extension to neural network-based uplift modeling, which is increasingly adopted in industry on observational and experimental data. By bridging statistical theory with machine learning practice, the paper aims to provide a structured overview for researchers and practitioners interested in causal effect estimation using modern computational tools.

1 Introduction

Causal inference, a hot topic in modern statistics, aims to estimate the causal effects of treatments beyond mere correlation. With the increasing availability of high-dimensional data and computational resources, machine learning methods have become essential tools in causal inference, especially for estimating heterogeneous treatment effect. This paper mainly focuses on and summarizes **the application of machine learning techniques within the framework of causal inference**.

Actually, the trial of combining Causal Inference with Machine Learning is a subtopic of how to combine the idea of Statistics with Computer Science, achieving the alignment of theory and practice. (So the topic this paper discussed can be attractive for statistics students who are interested in programming and also computer science students who are interested in math, and also non-math/cs students who are interested in both.)

> Q: Why titled *using Machine Learning for Causal Inference* instead of *Causal ML*?

It is necessary to clarify ahead that the term *Causal ML* may sometimes lead to confusion. This paper focuses on how to apply Machine Learning within the realm of Causal Inference. Which means that, the problems are causal, and the method is ML-based(see table 3 for detail). However, the term can also mean that the performance of ML models themselves can be improved with causality intuition[23][44]. So this paper is titled as the discussion is about the inference of causal estimands, rather than the performance of a prediction task.

This article is structured as follows. Chapter 1 is the introduction. In Chapter 2, I'll introduce the assumptions and estimands in causal inference, as well as some canonical theories, which importantly motivate the following discussion on ML-based methods. Chapter 3 focuses on the methodology of

Preprint. Under review.

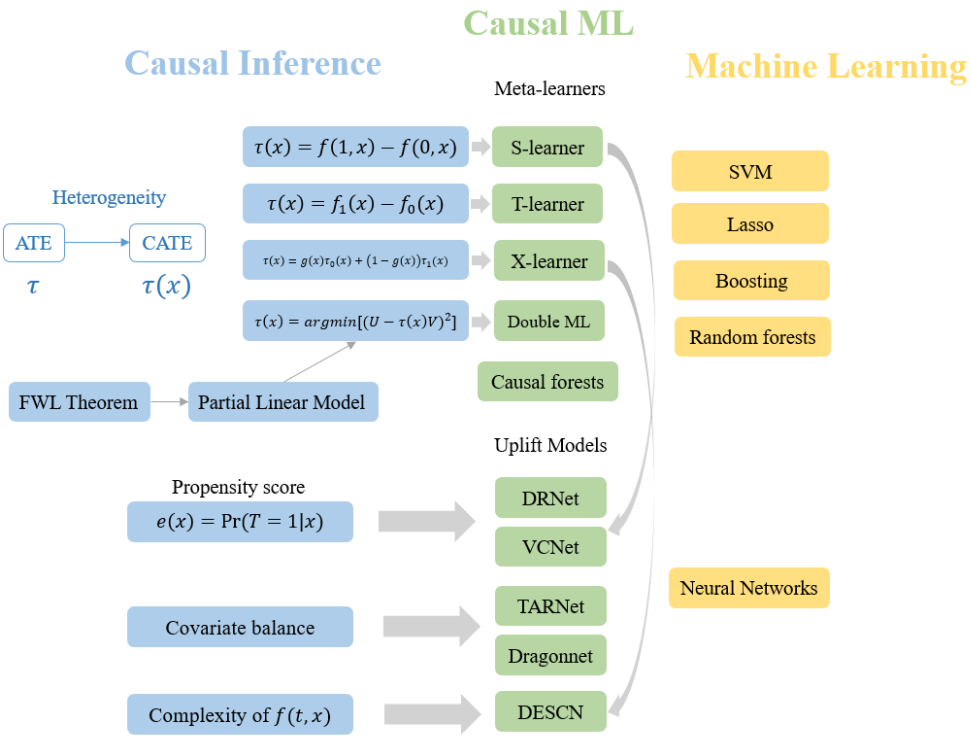

Figure 1: Roadmap

applying machine learning methods in causal scenarios, including meta-learners, tree-based causal forests and double machine learning. Chapter 4 details a recently developing direction in industry, which uses neural network-based uplift model to solve the high-dimensional problems with large dataset. Chapter 5 lists several widespread open-source implementations, books and tutorials as well as datasets.

The relationship of main chapters is summarized as follows: Firsly, a large part of methods in Chapters 3 and 4 are closely related to or intuited by theories in Chapter 2. So the word "recall" are used many times in this paper, indicating the connections of ML methods design with previous stated Causal statistical intuitions and equations. Additionally, the structures of chapter 4 can be seemed as an extension of the meta-learners introduced in chapter 3, which extends the methodology to Deep Networks and focuses mainly on high-dimensional, large dataset industrial applications. The overall road map or framework of this paper is displayed in figure 1.

2 What is Causal Inference

Before we dive in the question of how to combine causal inference with machine learning, it is important to clarify some basic concepts in causal inference. As a field in Statistic, it might be an unfamiliar topic for computer science researchers.

There are two commonly used frameworks of causal inference in statistics: the potential outcome framework(also known as the counterfactual framework)[34][18] and the causal diagram framework[29]. In this paper, we mainly focus on the potential outcome framework.

To simplify the writing, the following table1 gives the acronyms used in this paper as well as their full names.

Table 1: Acronyms

acronym	full name
SUTVA	Stable Unit Treatment Value Assumption
ATE	Average Treatment Effect
CATE	Conditional Average Treatment Effect
ML	Machine Learning
RCT	Randomized Controlled Trial
FWL	Frisch-Waugh-Lovell Theorem
DML	Double/Debiasd Machine Learning
NN	Neural Network

2.1 Basic set-ups, Assumptions and Estimands

Notations[34][14]: (1) Target Population: a well-defined population of units whose outcomes are going to be compared; (2) Treatment: an intervention, the effects of which the investigator wishes to assess relative to no intervention(i.e., the control). For simplicity, we consider a treatment with two levels: $t = 0, 1$; (3) Potential outcomes: the values of a unit's outcome (hypothetically) under treatment or control, $Y_i(0), Y_i(1)$.

Assumptions: (1) SUTVA[36][11] (2) Unconfoundedness[35] (3) Overlap[33]

The main estimand in causal inference is the **Average Treatment Effect(ATE)**, which is the average comparison between the potential outcomes under treatment and under control for the target population. There are also other estimands(see [14] problem 2.2 for details), such as Average treatment effect for the treated (ATT) or for the control (ATC), causal odds ratio (OR) and the Conditional Average Treatment Effect (CATE) which is discussed in detail in the next section.

$$\text{ATE} = \tau = \mathbb{E}[Y(1) - Y(0)] \tag{1}$$

Under the assumptions above, we have $Pr(Y(t)|X) = Pr(Y|X, T = t)$, so we can identify ATE as

$$\tau = E\{\mu_1(X) - \mu_0(X)\} = E\left\{\frac{TY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right\} \tag{2}$$

where $\mu_t(X) = E(Y(t)|X) = E(Y|T = t, X)$ is the outcome model under treatment $t(t = 0, 1)$.

2.2 From ATE to CATE: Treatment Effect Heterogeneity

Besides the ATE we discussed before, there is huge interest in understanding whether a treatment or policy affects certain individuals more than others[47][19][6], which is referred to as treatment effect heterogeneity or heterogeneous treatment effects, which leads to the estimation of the target estimand called Conditional Average Treatment Effect (CATE):

$$\text{CATE} = \tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] \tag{3}$$

Here is exactly where machine learning comes into play as an important tool, because the estimation of CATE coincides precisely with the industry's needs and technological advantages, as a large part of the questions in industry focus on precise treatment/action assignment, and ML-based method has an advantage in high-dimensional settings. Table2 lists some real world application examples, as well as their corresponding treatments, causal effects of interests.

Under assumptions above, we have $\mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}[Y \mid T = 1, X = x] - \mathbb{E}[Y \mid T = 0, X = x]$. Natually, two plug-in estimator of CATE can be constructed based on the following equations, which further inspire the s-learner and t-learner in section 3.1

$$\begin{aligned} f(t, x) &= \mathbb{E}[Y \mid T = t, X = x] \\ \tau(x) &= f(1, x) - f(0, x) \end{aligned} \tag{4}$$

$$\begin{aligned} f_1(x) &= \mathbb{E}[Y \mid T = 1, X = x] \\ f_0(x) &= \mathbb{E}[Y \mid T = 0, X = x] \\ \tau(x) &= f_1(x) - f_0(x) \end{aligned} \tag{5}$$

Table 2: Examples of Industrial Scenarios with Heterogeneous Treatment Effects

Industry Scenario	Treatment	Causal Effect of Interest
E-commerce[47][20]	whether showing a coupon or not to user	the effect of coupon on the user's purchase amount
Online advertising[6][24]	whether showing an ad to a user	the effect of ad exposure on click-through rate or conversion
Precision medicine[19]	whether the patient take the medicine or not	the effect of the medicine on the patient's recovery level

2.3 Two important thing for observational data: covariate balance and propensity score

In observational studies, in contrast with Randomized Controlled Trial(RCT), we do not know the assignment mechanism. Ensuring covariate overlap and balance is the key for outcome modeling in causal inference[41][37], as if the imbalance is severe, the model-based results heavily relies on the region with little overlap, which makes overfitting a concern[32]. The most common metric of balance is the absolute standardized difference (ASD).

$$\text{ASD} = \left| \frac{\sum_{i=1}^{N} X_i Z_i}{\sum_{i=1}^{N} Z_i} - \frac{\sum_{i=1}^{N} X_i (1 - Z_i)}{\sum_{i=1}^{N} (1 - Z_i)} \right| \Bigg/ \sqrt{\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}}, \tag{6}$$

Propensity score[33], defined as the conditional probability of receiving a treatment given covariates X, is an importance thing in causal inference, as a summary statistic of the assignment mechanism.

$$e(x) = Pr(T = 1|x) = E(T|X) \tag{7}$$

The good property of propensity score is that, it balances the distribution of covariates between treatment groups. Additionally, it can be viewed as a summary score of the observed covariates as if T is unconfounded given x, then it is unconfounded given $e(X)$.

$$T \perp X \mid e(X) \tag{8}$$

$$\{Y_i(1), Y_i(0)\} \perp T_i \mid X_i \implies \{Y_i(1), Y_i(0)\} \perp T_i \mid e(X_i) \tag{9}$$

These insights have guided the handling of the treatment variable t in the design framework of the uplift models in chapter 4.

2.4 FWL and Robinson decomposition

Frisch-Waugh-Lovell(FWL) Theorem[16][25] is a classic theorem in economics. Instead of OLS, it gives another consistent way to estimate a coefficient in a linear model by regressing the residuals.

Theorem 1 (FWL Theorem) *In OLS regression $Y = X\hat{\beta} + \hat{\varepsilon}$, if we partition X and β into*

$$X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

where $X_1 \in \mathbb{R}^{n \times k}, X_2 \in \mathbb{R}^{n \times l}, \beta_1 \in \mathbb{R}^k$ and $\beta_2 \in \mathbb{R}^l$, then we have

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon},$$

The OLS estimator for β_2 in the regression has the equivalent form

$$\begin{aligned} \hat{\beta}_2 &= \left\{X_2^\top (I_n - H_1) X_2\right\}^{-1} X_2^\top (I_n - H_1) Y \quad \textit{where } H_1 = X_1 (X_1^\top X_1)^{-1} X_1^\top \\ &= (\tilde{X}_2^\top \tilde{X}_2)^{-1} \tilde{X}_2^\top \tilde{Y} \quad \textit{where } \tilde{X}_2 = (I_n - H_1) X_2 \textit{ and } \tilde{Y} = (I_n - H_1) Y. \end{aligned}$$

Robinson [31] generalized FWL theorem by replacing the linear regression with some nonparametric (e.g. kernel) regression.

Theorem 2 (Generalized Partial Linear Model)

$$\begin{gathered} Y = T\tau(X) + m(X) + U, \quad \mathbb{E}(U \mid X, T) = 0 \\ T = e(X) + V, \quad \mathbb{E}(V \mid X) = 0 \\ \textit{with } E(UV|X,T) = 0, \textit{ where } \tau(X) \textit{ is the CATE} \end{gathered} \tag{10}$$

Same idea with FWL theorem, we can estimate $\tau(x)$ by regressing residualized outcome $\hat{U}$ and treatment $\hat{V}$

$$\hat{\tau}(X) = \arg\min_{\tau \in \mathcal{E}_n} \left[(\hat{U} - \tau(X)\hat{V})^2 \right] \tag{11}$$

which is further generalized to machine learning models as Double Machine Learning(DML), which is a popular way of using ML in causal inference, and details be discussed in section 3.3.

3 ML-based Methods for Causal Inference

There has been a dramatic increase in semiparametric or nonparametric estimators of the CATE that utilize modern statistical learning tools. Let's think about a question firstly: ML methods are designed for prediction, how about causal counterfactual prediction task? What's the difference between traditional ML task and causal task?

Q: The difference between traditional ML task and causal ML task?

Table 3: Traditional ML task v.s Causal ML task

Class	Data	Estimands	Evaluation
Traditional ML task	(x_i, Y_i)	$E(Y_i \mid X = x_i)$	mse/mae/auc,...
Causal ML task	$(x_i, t_i, Y_i(t))$	a function of $Y_i(1), Y_i(0)$ e.g. $E(Y_i(1) - Y_i(0) \mid X = x_i)$	auuc/bam/eom,...

As stated in table 3, the data, estimands, as well as the evaluation procedure, are all different in causal tasks compared with traditional ML prediction tasks. Therefore, one could not simply applying ML method in causal inference, because the problem can not be transformed into a prediction manner task. So several approaches are developed to bridge these two, making it possible to utilize powerful ML learners in causal estimations.

In general, ML-based methods play an important role in estimating causal effects, which is summarized into meta-learners, causal forests and double machine learning discussed in this chapter.

3.1 Meta-learners

3.1.1 S-learner and T-learner: Estimating $f(t, x)$ and $f_t(x)$

Recall the method of CATE indentification(e.q. 4), one way of estimating the CATE is by estimates of the outcome model. This idea leads to S-learners(S refers to Single), which focus solely on flexible estimation of $f(t, x) = E[Y|X = x, T = t]$ and CATE estimation is automatic after this.

$$\hat{\tau}(x) = \hat{f}(1, x) - \hat{f}(0, x) \tag{12}$$

The simplest approach is with a linear model, and Machine learning (trees, regularized regressions, boosting, etc.) can also be used. One of the seminal papers is the Bayesian Additive Regression Trees(BART)[1][10], which places certain priors on the parameters of the tree, and regression trees[4].

One important point for s-learner is whether to regard the treatment variable as of the same status as the other variables. If the dataset we get is high-dimensional, which is always the case in industrial scenarios, then directly putting in t with x would lead to submerged effect of the treatment in the deep model[27][38][47]. More will be discussed in section 4.

A very related approach is to specify separate models in the treated and control groups, which is called T-learners(T refers to Two). Recall equation 5, the plug-in estimator for estimating CATE need to fit the outcome model in each treatment group:

$$\hat{\tau}(x) = \hat{f}_1(x) - \hat{f}_0(x) \tag{13}$$

3.1.2 X-learner: Estimating $f_t(x)$ with $\tau(x)$

S-learners and T-learners seems direct and mathematically-founded, and one can use any supervised ML method(e.g., Lasso, random forests, neural networks) as base learner and estimate CATE using plug-in estimation of $\hat{f}(t,z), \hat{f}_t(x)$. However, ML-based model can be highly nonlinear and fairly complex, flexible, so we should be careful about the implications of our prior specification on the parameter of interest. Let's think about a question:

> Q: Do we expect the CATE to be as complex as $f(t,x)$?

This actually talks about what if the relation between Y and X is more complex than Y and T? This makes sense, and there is case when the individual regression functions are complex, but the overall difference between them is zero[21].

This enlightened the design of X-learner(X refers to Cross)[22], which tackles the over-complex-problem, by using information from the control group to derive better estimators for the treatment group and vice versa. By leveraging the structural details of the CATE, it empowers us to handle the complexity problem and effectively exploit an unbalanced design. In X-learner, CATE is estimated by a weighted average of two models, combining the information of both treatment groups. This also inspired the X-network in section 4.

3.2 Causal Forests: Directly Estimating $\tau(x)$

Causal trees and forests[43][5][30] are effective for causal task, because it directly estimates the CATE by finding subsets of the data where the treatment effect varies the most. Different from the random forests[8], which pick splits that reduce the MSE(Or Gini index / classification error for categorical outcomes) the most among all possible splits, in causal trees, we split to maximize heterogeneity of the treatment effect. Recall that CATE is just about Heterogeneity of treatment effect, so that the procedure is direct.

$$\frac{|\widehat{\tau}_l - \widehat{\tau}_r|}{\sqrt{\widehat{\text{Var}}(\widehat{\tau}_l) + \widehat{\text{Var}}(\widehat{\tau}_r)}} \tag{14}$$

In random forests, we use cross-validation to tune hyper parameters[8], but this procedure is not that natural for causal scenarios, as calculating MSE is impossible. Therefore, Honesty criterion[3][43] is developed for tuning a causal tree model, as a sample can only be used to estimate CATE or decide how to build the model, but not both. By adapting Honesty criterion and divide the sample into three parts, we can tuning a causal tree without the trouble of overfitting.

3.3 Double Machine Learning: Estimating $\tau(x)$ by regressing residuals

A recurring idea is "double" learning: using ML for both outcome and propensity model, and combine, in which a general theoretical framework is Double/Debiased Machine Learning by Chernozhukov et al.[9]. Recall the FWL and generalized partial linear model(e.q. 10) in section 2.4, which implies that we can identify CATE by regressing residuals. DML[9] uses these ideas and designs a three-step manner, which enables the use of machine learning models in a 'double' manner for causal inference.

$$\tau(\cdot) = \arg\min_{\tau} \left\{ \mathbb{E}\left[((Y_i - m(X_i)) - (T_i - e(X_i))\tau(X_i))^2 \right] \right\} \tag{15}$$

There are different choices of $\tau(x)$ in DML for CATE, like Reproducing Kernel Hilbert Space[28], Random forest[5], and Sparse linear space[9]. Besides, DML also coincides with a parameterization[31] idea which directly estimates the CATE by minimizing the loss function above, which is called R-learner[28].

4 Industrial applications and Advanced Topics: NN-based Uplift Models

Recall the applications in table2 and what we discuss in section 2.2, based on the large dataset nowadays and the convenience of conducting Randomized Controlled Trial(RCT) in mobile platforms we use everyday, NN-based model becomes possible. The term *Uplift Model*("uplift" refers precisely to the treatment effect) is becoming increasingly common in internet companies, which is mainly about using deep models for the estimating of CATE. And this field has found wide practical applications in the industry[45][46]. The following are several recent structures of uplift models, each of which reflects the corresponding statistical intuition.

Q: How do statistical properties inspire the design of NN-based model structure?

Continuous treatment Continuous treatment situations is frequently encountered in reality, although they encounter inherent difficulty in theory like poor overlap [7]. Dose Response Network (DRNet)[38] divides the treatment into several equal intervals, with each branch learning the corresponding result, to approximately estimate the causal effect of the continuous treatment. By discretizing the continuous treatment, it can be applied to complex scenarios like multi-treatment or continuous treatments. However, this piecewise regression method leads to the problem of sudden changes at the boundary points of the segmented intervals. Varying Coefficient Neural Network (VCNet)[27] can address this issue by leveraging spline functions to constrain the derivatives at the breakpoints, thereby ensuring continuity across these points. Additionally, the targeted regularization method is generalized to obtain a doubly robust estimator.

NN-based model for observational study Recall the importance of covariate balance in observational study in section 2.3, the different distribution of covariates in the treated and control groups make the estimation unstable and rely much on the model specification. Treatment-Agnostic Representation Network (TARNet)[39], which tries to learn a "balanced" representation, exactly coincides with this idea. Also, recall the importance of propensity score for observational data in section 2.3. In order to accurately predict the CATE instead of focusing on the prediction of the outcome Y, Dragonnet[40], as an extension of TARNet[39], imposes an additional structure which extracts the parts of X that are related to T. By modeling propensity score and conditional outcome simultaneously, the model can achieve a better performance in observational studies. A regularization procedure based on non-parametric estimation theory is also proposed.

X-Network Inspired by the x-learner[22] method in section 3.1.2, Deep Entire Space Cross Networks (DESCN)[47] tackles the treatment bias issue through the Entire Space Network (ESN) and mitigates the sample imbalance problem by using the Cross Network, which jointly learns the true responses, and the pseudo treatment effect in an integrated manner. By utilizing the information as well as the relation of outcome in each treatment groups crossly, it improves the performance of CATE greatly. This structure, in some sense, reduces the complexity of the outcome model, which enables stable estimation, similar idea is also applied in FlexTENet[12].

Once we have the question modeled into a deep learning structure, one can then use any classic Network design techniques in Deep Learning, such as Local Hidden Unit Contribution (LHUC)[42] and Mixture of Experts (MMoE)[26], to further improve the representational performance of the model or tackle multi-task problem.

Table 4: NN-based Uplift Models

Name	Main point
DRNet[38]	discretizing the continuous treatment into branches
VCNet[27]	using spline functions for continuous treatment
TARNet[39]	balance the covariate distribution in treated and control groups
Dragonnet[40]	extracts the parts of X that are important for propensity score
DESCN[47]	x-network which utilizing the pseudo treatment effect
FlexTENet[12]	multiple ways of control the complexity of estimation

5 Resources

For the topic discussed, some open-source implementations, including python package and R library, books and tutorials, as well as public datasets(more can be found in https://github.com/gsbDBI/ExperimentData), are shown in the following tables.

Table 5: Open-Source Implementations

Name	Description	url
EconML	Python package	https://econml.azurewebsites.net/
CausalML	Python package	https://causalml.readthedocs.io/en/latest/about.html
DoubleML	R package	https://docs.doubleml.org/stable/index.html
PSweight	R package	https://cran.r-project.org/web/packages/PSweight/index.html

Table 6: Books and Tutorials

Author	Name	url
Hernán MA, Robins JM	Causal Inference: what if [17]	https://miguelhernan.org/whatifbook
Peng Ding	A First Course in Causal Inference[14]	https://arxiv.org/abs/2305.18793
Fan Li	STA 640: Causal Inference	https://www2.stat.duke.edu/~fl35/CausalInferenceClass.html
Susan Athey	MGTECON 634: Machine Learning and Causal Inference	https://explorecourses.stanford.edu/search?q=MGTECON%20634&academicYear=20212022

Table 7: Datasets

Name	url
Criteo Uplift Modeling Dataset[13]	https://ailab.criteo.com/criteo-uplift-prediction-dataset/
Epileptic Seizure Recognition Dataset[2]	https://www.ukbonn.de/epileptologie/arbeitsgruppen/ag-lehnertz-neurophysik/downloads/
AirBnB Dataset[15]	https://github.com/gsbDBI/ExperimentData/tree/master/AirBnB
Medical dataset	https://github.com/higgi13425/medicaldata/

References

[1] Jennifer L. Hill and. "Bayesian Nonparametric Modeling for Causal Inference". In: *Journal of Computational and Graphical Statistics* 20.1 (2011), pp. 217–240. DOI: 10.1198/jcgs.2010.08162. eprint: https://doi.org/10.1198/jcgs.2010.08162. URL: https://doi.org/10.1198/jcgs.2010.08162.

[2] Ralph G Andrzejak et al. "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state". In: *Physical Review E* 64.6 (2001), p. 061907.

[3] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7353–7360.

[4] Susan Athey and Guido W Imbens. "Machine learning methods for estimating heterogeneous causal effects". In: *stat* 1050.5 (2015), pp. 1–26.

[5] Susan Athey, Julie Tibshirani, and Stefan Wager. *Generalized Random Forests*. 2018. arXiv: 1610.01271 [stat.ME]. URL: https://arxiv.org/abs/1610.01271.

[6] Marianne Bertrand et al. "What's advertising content worth? Evidence from a consumer credit marketing field experiment". In: *The quarterly journal of economics* 125.1 (2010), pp. 263–306.

[7] Michela Bia and Alessandra Mattei. "A Stata package for the estimation of the dose-response function through adjustment for the generalized propensity score". In: *The Stata Journal* 8.3 (2008), pp. 354–373.

[8] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[9] Victor Chernozhukov et al. *Double/Debiased Machine Learning for Treatment and Causal Parameters*. 2024. arXiv: 1608.00060 [stat.ML]. URL: https://arxiv.org/abs/1608.00060.

[10] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1 (Mar. 2010). ISSN: 1932-6157. DOI: 10.1214/09-aoas285. URL: http://dx.doi.org/10.1214/09-AOAS285.

[11] Stephen R Cole and Constantine E Frangakis. "The consistency statement in causal inference: a definition or an assumption?" In: *Epidemiology* 20.1 (2009), pp. 3–5.

[12] Alicia Curth and Mihaela Van der Schaar. "On inductive biases for heterogeneous treatment effect estimation". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15883–15894.

[13] Diemert Eustache, Betlei Artem, Christophe Renaudin, and Amini Massih-Reza. "A Large Scale Benchmark for Uplift Modeling". In: *Proceedings of the AdKDD and TargetAd Workshop, KDD, London,United Kingdom, August, 20, 2018*. ACM, 2018.

[14] Peng Ding. *A First Course in Causal Inference*. 2023. arXiv: 2305.18793 [stat.ME]. URL: https://arxiv.org/abs/2305.18793.

[15] Benjamin Edelman, Michael Luca, and Dan Svirsky. "Racial discrimination in the sharing economy: Evidence from a field experiment". In: *American economic journal: applied economics* 9.2 (2017), pp. 1–22.

[16] Ragnar Frisch and Frederick V Waugh. "Partial time regressions as compared with individual trends". In: *Econometrica: Journal of the Econometric Society* (1933), pp. 387–401.

[17] Miguel A Hernán and James M Robins. *Causal inference*. 2010.

[18] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[19] Maciej Jaskowski and Szymon Jaroszewicz. "Uplift modeling for clinical trial data". In: *ICML workshop on clinical data analysis*. Vol. 46. 2012, pp. 79–95.

[20] Wenwei Ke et al. "Addressing exposure bias in uplift modeling for large-scale online advertising". In: *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2021, pp. 1156–1161.

[21] Edward H. Kennedy. *Towards optimal doubly robust estimation of heterogeneous causal effects*. 2023. arXiv: 2004.14497 [math.ST]. URL: https://arxiv.org/abs/2004.14497.

[22] Sören R Künzel et al. "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10 (2019), pp. 4156–4165.

[23] Haoxuan Li et al. "Propensity matters: measuring and enhancing balancing for recommendation". In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023.

[24] Dugang Liu et al. *Explicit Feature Interaction-aware Uplift Network for Online Marketing*. 2023. arXiv: 2306.00315 [cs.LG]. URL: https://arxiv.org/abs/2306.00315.

[25] Michael C Lovell. "Seasonal adjustment of economic time series and multiple regression analysis". In: *Journal of the American Statistical Association* 58.304 (1963), pp. 993–1010.

[26] Jiaqi Ma et al. "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts". In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 1930–1939.

[27] Lizhen Nie et al. *VCNet and Functional Targeted Regularization For Learning Causal Effects of Continuous Treatments*. 2021. arXiv: 2103.07861 [cs.LG]. URL: https://arxiv.org/abs/2103.07861.

[28] Xinkun Nie and Stefan Wager. "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2 (2021), pp. 299–319.

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

[30] Scott Powers et al. "Some methods for heterogeneous treatment effect estimation in high dimensions". In: *Statistics in medicine* 37.11 (2018), pp. 1767–1787.

[31] Peter M Robinson. "Root-N-consistent semiparametric regression". In: *Econometrica: Journal of the Econometric Society* (1988), pp. 931–954.

[32] Paul R Rosenbaum. "Model-based direct adjustment". In: *Journal of the American statistical Association* 82.398 (1987), pp. 387–394.

[33] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55.

[34] Donald B Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[35] Donald B Rubin. "Inference and missing data". In: *Biometrika* 63.3 (1976), pp. 581–592.

[36] Donald B Rubin. "Randomization analysis of experimental data: The Fisher randomization test comment". In: *Journal of the American statistical association* 75.371 (1980), pp. 591–593.

[37] Donald B Rubin. "The use of matched sampling and regression adjustment to remove bias in observational studies". In: *Biometrics* (1973), pp. 185–203.

[38] Patrick Schwab et al. "Learning Counterfactual Representations for Estimating Individual Dose-Response Curves". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04 (Apr. 2020), pp. 5612–5619. ISSN: 2159-5399. DOI: 10.1609/aaai.v34i04.6014. URL: http://dx.doi.org/10.1609/aaai.v34i04.6014.

[39] Uri Shalit, Fredrik D. Johansson, and David Sontag. *Estimating individual treatment effect: generalization bounds and algorithms*. 2017. arXiv: 1606.03976 [stat.ML]. URL: https://arxiv.org/abs/1606.03976.

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

[41] Elizabeth A. Stuart. "Matching Methods for Causal Inference: A Review and a Look Forward". In: *Statistical Science* 25.1 (Feb. 2010). ISSN: 0883-4237. DOI: 10.1214/09-sts313. URL: http://dx.doi.org/10.1214/09-STS313.

[42] Pawel Swietojanski, Jinyu Li, and Steve Renals. "Learning hidden unit contributions for unsupervised acoustic model adaptation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.8 (2016), pp. 1450–1463.

[43] Stefan Wager and Susan Athey. *Estimation and Inference of Heterogeneous Treatment Effects using Random Forests*. 2017. arXiv: 1510.04342 [stat.ME]. URL: https://arxiv.org/abs/1510.04342.

[44] Peng Wu et al. *On the Opportunity of Causal Learning in Recommendation Systems: Foundation, Estimation, Prediction and Challenges*. 2022. arXiv: 2201.06716 [cs.IR]. URL: https://arxiv.org/abs/2201.06716.

[45] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. "GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=ByKWUeWA-.

[46] Yi-Fan Zhang et al. *Exploring Transformer Backbones for Heterogeneous Treatment Effect Estimation*. 2022. arXiv: 2202.01336 [cs.LG]. URL: https://arxiv.org/abs/2202.01336.

[47] Kailiang Zhong et al. "DESCN: Deep Entire Space Cross Networks for Individual Treatment Effect Estimation". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '22. Washington DC, USA: Association for Computing Machinery, 2022, pp. 4612–4620. ISBN: 9781450393850. DOI: 10.1145/3534678.3539198. URL: https://doi.org/10.1145/3534678.3539198.