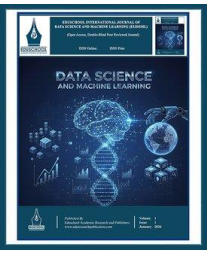
EDUSCHOOL INTERNATIONAL JOURNAL OF DATA SCIENCE AND MACHINE LEARNING (EIJDSML)

(Open Access, Double-Blind Peer Reviewed Journal)

ISSN Online:



Causal Inference and Machine Learning for Reliable Decision-Making

Rasmi M

Assistant Professor, Department of Computer Applications, Bethania Institute of Management Studies, Kunnamkulam, Kerala, India

Article information

Received: 12th March 2026
Received in revised form: 15th April 2026
Accepted: 18th May 2026
Available online: 18th July 2026

Volume: 1
Issue: 2
DOI: https://doi.org/10.5281/zenodo.21410727

Abstract

Machine learning systems trained to predict outcomes from observational data often fail when their outputs guide interventions. A model that accurately forecasts who will churn, default, or recover does not by itself reveal what happens if a policy changes the treatment assigned to those people. This gap between prediction and intervention is the central concern of causal inference. This paper reviews the methodological bridge connecting causal reasoning with modern supervised learning and explains why correlation-driven models give biased answers to questions about actions. The potential outcomes framework and structural causal models are introduced as complementary languages for stating the target of estimation. Identification is treated through directed acyclic graphs, the backdoor criterion, and instrumental variables, which specify when an effect can be recovered from data. The discussion then turns to estimation of average and conditional average treatment effects using meta-learners (S, T, and X), causal forests, and double or debiased machine learning, together with uplift modeling for targeting decisions. A synthetic benchmark illustrates typical error behavior: orthogonalized and forest-based estimators reach an absolute ATE bias near 0.05 and lower CATE error than a single-model baseline. Applications in pricing, marketing, healthcare, and public policy are surveyed, along with validation difficulties that arise because counterfactual outcomes are never observed. The paper closes with the assumptions, chiefly unconfoundedness and overlap, on which every reported number depends.

Keywords:- Causal Inference, Treatment Effect Estimation, Double Machine Learning, Causal Forests, Uplift Modeling, Confounding, Potential Outcomes.

.

I. INTRODUCTION

Supervised learning answers questions of the form: given features X, what value of the outcome Y should be expected? Such predictions support many useful tasks, from ranking search results to flagging fraudulent transactions. A different class of questions drives most operational decisions. A clinician asks whether a drug will reduce a patient's risk, not merely whether high-risk patients tend to receive it. A marketing team asks whether a discount causes a purchase, not whether discount recipients buy more. These are questions about interventions, and they cannot be settled by accuracy on held-out data [1], [2]. The difficulty is structural. Observational datasets record decisions that were already made, often by agents who acted on information correlated with the outcome. Patients in poor health receive aggressive treatment; price-sensitive shoppers receive coupons. A predictive model fitted to such data learns the joint behavior of the assignment process and the outcome process, and it cannot separate them without additional assumptions [3]. When the assignment mechanism shifts, as it does whenever a

new policy is deployed, the model's correlational structure no longer holds. Causal inference supplies the missing structure. It defines the estimand, the quantity a decision maker actually wants, and states the conditions under which that quantity can be recovered from data. Over the past decade a body of work has joined this classical theory with flexible machine learning, producing estimators that use neural networks, gradient boosting, and random forests as components while retaining valid statistical guarantees [4], [5], [6]. This paper reviews that synthesis. Section II presents the potential outcomes framework and structural causal models. Section III treats identification and confounding. Section IV reviews ML estimators for treatment effects. Section V reports a synthetic benchmark and discusses applications. Section VI examines assumptions and limitations, and Section VII concludes.

II. FOUNDATIONS: POTENTIAL OUTCOMES AND STRUCTURAL CAUSAL MODELS

A. Potential Outcomes

The Neyman-Rubin potential outcomes model assigns to each unit i two outcomes: $Y_i(1)$, the outcome under treatment, and $Y_i(0)$, the outcome under control [1], [7]. The individual treatment effect is the difference $Y_i(1)-Y_i(0)$. Only one of the two potential outcomes is ever observed, since a unit is either treated or not, a feature Holland termed the fundamental problem of causal inference [8]. The observed outcome relates to the potential outcomes through the assignment indicator:

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0) \quad (1)$$

Because individual effects are unidentifiable without strong assumptions, attention centers on averages. The average treatment effect (ATE) is the expectation of Y(1) - Y(0) over the population. The conditional average treatment effect (CATE), written:

$$\tau(x) = E[Y(1) - Y(0) \mid X = x] \quad (2)$$

This describes how the effect varies with covariates and is the target of personalized decision rules [4], [9]. A valid interpretation of any estimate rests on the stable unit treatment value assumption, which rules out interference between units and multiple versions of treatment [7].

B. Structural Causal Models and DAGs

Pearl's structural causal model represents a system as a set of variables and functions, each variable computed from its direct causes and an independent noise term [2], [10]. The qualitative skeleton of such a model is a directed acyclic graph (DAG), in which an arrow from A to B states that A appears in the function generating B. The do-operator formalizes intervention: do(T = t) replaces the function for T with the constant t, severing the arrows into T while leaving the rest of the system intact. The interventional distribution P(Y | do(t)) is generally different from the conditional distribution P(Y | t), and the difference is exactly what confounding measures. Fig. 1 shows a compact model with a confounder X that influences both treatment and outcome, a mediator M on the path from treatment to outcome, and an unobserved factor U. The graph makes the estimand and its obstacles visible at a glance: the total effect of T on Y travels along the direct edge and through M, while the backdoor path through X, and any path through U, must be blocked for the effect to be identified.

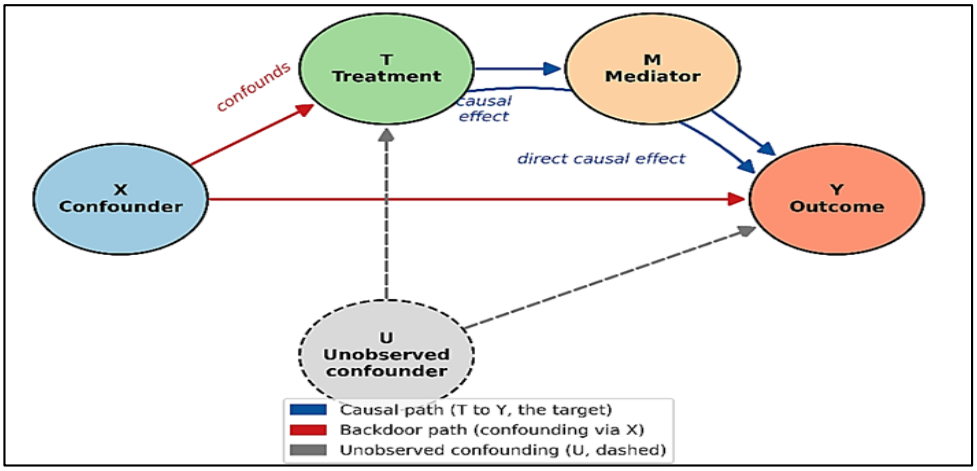


Fig. 1: A structural causal graph with treatment T, confounder X, mediator M, outcome Y, and unobserved factor U. Blue marks the causal path the analyst wants; red marks the backdoor path through X that must be blocked; dashed grey marks unobserved confounding. Takeaway: the graph names exactly what to estimate and what to adjust away.

III. IDENTIFICATION AND CONFOUNDING

Identification asks whether a causal quantity can be written as a function of the observed distribution. It is a property of the assumed model, not of the sample size, and no amount of data repairs a non-identified estimand. The clearest illustration is confounding. When a common cause influences both treatment and outcome, the naive difference in observed means mixes the treatment effect with the confounder's imprint.

Fig. 2 demonstrates the reversal that confounding can produce. In the simulated cohort, disease severity drives both the decision to treat and the outcome. Comparing treated and untreated patients directly suggests that treatment worsens the outcome, because sicker patients are treated more often. Stratifying on severity reveals the true effect, which is protective. This sign reversal is a finite-sample analogue of Simpson's paradox and a reminder that adjustment, not raw correlation, carries the causal content [11].

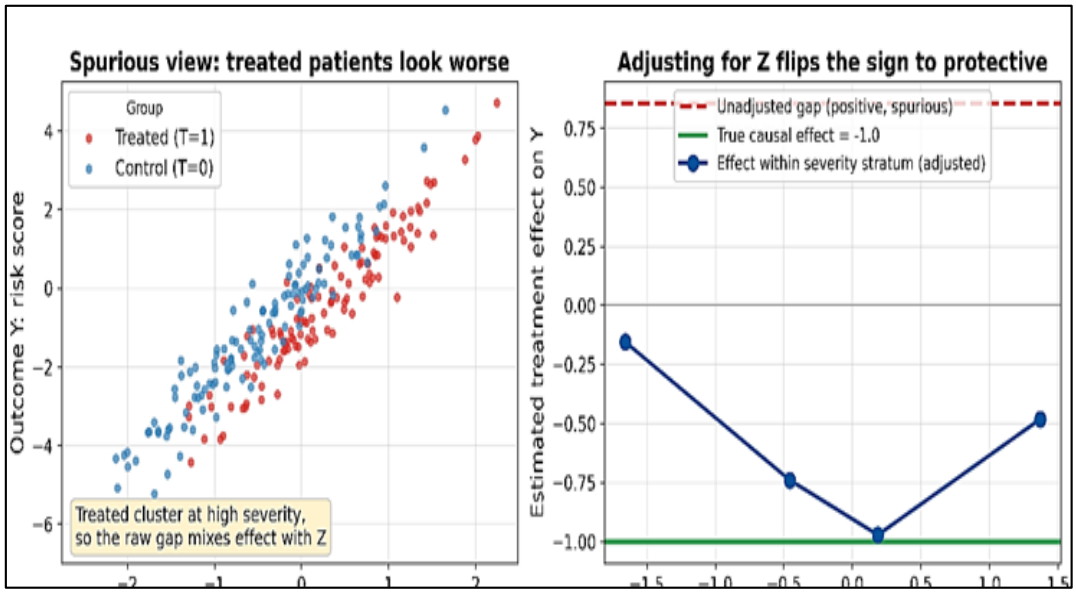


Fig. 2: A confounder reverses an apparent association. The unadjusted comparison (left) suggests harm because treated patients are sicker; stratifying on severity (right) recovers the protective effect. Takeaway: the sign of an effect can flip once the confounder is held fixed.

A. The Backdoor Criterion

A set of covariates Z satisfies the backdoor criterion relative to (T, Y) if no node in Z is a descendant of T and Z blocks every path from T to Y that begins with an arrow into T [2]. When such a Z is observed, the interventional distribution is identified by adjustment: P(Y | do(t)) equals the average over Z of P(Y | t, Z). Conditioning on the right set removes confounding, while conditioning on a mediator or a collider can introduce bias, so the choice of adjustment set follows from the graph rather than from a search for predictive features [12].

B. Instrumental Variables

When an unobserved confounder remains, as with U in Fig. 1, backdoor adjustment is unavailable. An instrumental variable offers an alternative route [13]. An instrument affects treatment, is independent of the unobserved confounders, and influences the outcome only through treatment. Variation in the instrument induces variation in treatment that is free of confounding, and under a monotonicity condition it identifies a local average treatment effect for the subpopulation whose treatment responds to the instrument [14]. Encouragement designs, policy discontinuities, and randomized eligibility are common sources of instruments in applied work.

IV. MACHINE LEARNING ESTIMATORS FOR TREATMENT EFFECTS

Once an estimand is identified, the remaining task is estimation from finite data. Classical methods such as matching and inverse propensity weighting remain useful, but they struggle with many covariates and complex response surfaces. Machine learning supplies flexible function approximators, and recent estimators wrap those approximators in procedures that preserve statistical validity [4], [5], [6].

A. Meta-Learners

Meta-learners reduce CATE estimation to standard regression [9]. The S-learner trains one model on covariates and treatment jointly, then differences its predictions at T = 1 and T = 0. It is simple but tends to shrink the estimated effect toward zero when the base learner regularizes the treatment indicator. The T-learner fits separate models for treated and control groups, which avoids that bias but wastes data when one group is small. The X-learner adds a cross step that imputes individual effects and reweights by the propensity score, improving accuracy under imbalanced assignment, a common situation in observational studies.

B. Causal Forests

Generalized random forests estimate heterogeneous effects by growing trees that split to maximize differences in treatment effect rather than differences in outcome [4], [15]. Honest sample splitting, in which one subsample chooses splits and another estimates effects within leaves, yields pointwise confidence intervals and asymptotic normality. Causal forests adapt to the covariates that drive heterogeneity and degrade gracefully when effects are uniform.

C. Double or Debiased Machine Learning

Double machine learning targets a low-dimensional effect while using arbitrary learners for the high-dimensional nuisance functions, namely the outcome regression and the propensity score [5]. Two devices protect the estimate. Neyman orthogonality makes the moment condition insensitive to small errors in the nuisance estimates, and cross-fitting removes the bias that arises from using the same data to fit nuisances and the effect. The result is a root-n consistent, asymptotically normal estimate of the ATE even when the nuisance learners converge slowly. Doubly robust scores combine the outcome model and the propensity model so that consistency holds if either is correct [16], [17].

D. Representation Learning and Deep Models

Neural approaches learn balanced representations in which treated and control distributions overlap. TARNet and counterfactual regression add an imbalance penalty between the two groups in representation space, reducing the error of individual effect estimates [18]. The causal effect variational autoencoder treats an unobserved confounder as a latent variable inferred from proxies [19]. These methods extend treatment effect estimation to images, text, and other unstructured covariates, at the cost of weaker theoretical guarantees than orthogonalized estimators provide.

Table 1. Machine learning estimators for treatment effects.

Estimator	Target	Base learner	Key property
S-learner	CATE	Any regressor	Simple; can bias effect toward zero
T-learner	CATE	Two regressors	Unbiased split; data-hungry
X-learner	CATE	Regressors + propensity	Strong under imbalance
Causal forest	CATE	Honest trees	Valid intervals; adapts to heterogeneity
Double ML	ATE	Any nuisance learner	Orthogonal; cross-fitted; root-n
TARNet / CFR	CATE	Neural network	Balanced representation; unstructured X

V. APPLICATIONS AND ILLUSTRATIVE RESULTS

A. A Synthetic Benchmark

To compare estimators on common ground, a synthetic dataset was generated with a known treatment effect, correlated covariates, confounded assignment, and heterogeneous response. Because the ground truth is fixed by construction, both ATE bias and the precision of conditional estimates can be measured directly. The figures reported here are illustrative of typical behavior and do not describe a deployed system. Fig. 3 summarizes absolute ATE bias and a CATE root mean squared error in the style of the PEHE metric across five estimators.

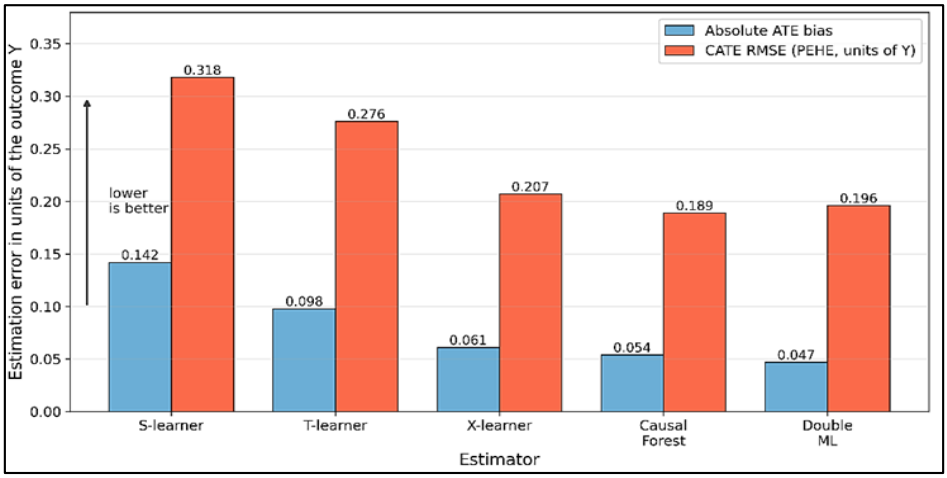


Fig. 3: Absolute ATE bias and CATE RMSE for five estimators on a synthetic benchmark with known ground truth. Lower bars are better. Takeaway: causal forests and double machine learning recover the true effect most accurately, while the S-learner is the most biased.

The pattern matches theory. The S-learner shows the largest bias because the base learner regularizes the treatment indicator. The X-learner improves markedly under the imbalanced assignment built into the simulation. Causal forests and double machine learning achieve the smallest ATE bias, near 0.05 in absolute terms, with the forest giving the lowest conditional error. Table 2 lists the numbers and adds an interval coverage column that reports how often nominal ninety-five percent intervals contain the truth.

Table 2. Synthetic benchmark results. Values are illustrative, not from a deployed study.

Estimator	|ATE bias|	CATE RMSE	95% interval coverage
S-learner	0.142	0.318	0.86
T-learner	0.098	0.276	0.89
X-learner	0.061	0.207	0.92
Causal forest	0.054	0.189	0.94
Double ML	0.047	0.196	0.95

B. Uplift Modeling for Targeting

Many business decisions need a ranking of who responds to an action, not a single average. Uplift modeling estimates the individual change in response probability and ranks the population by it [20], [21]. Targeting the high-uplift segment concentrates spending where treatment changes behavior, while sparing both the customers who would act anyway and those who react negatively. The Qini curve evaluates such rankings by plotting cumulative incremental response against the fraction of the population treated. Fig. 4 contrasts an uplift model and a causal forest with random targeting; the area between a model curve and the diagonal, the Qini coefficient, summarizes the gain.

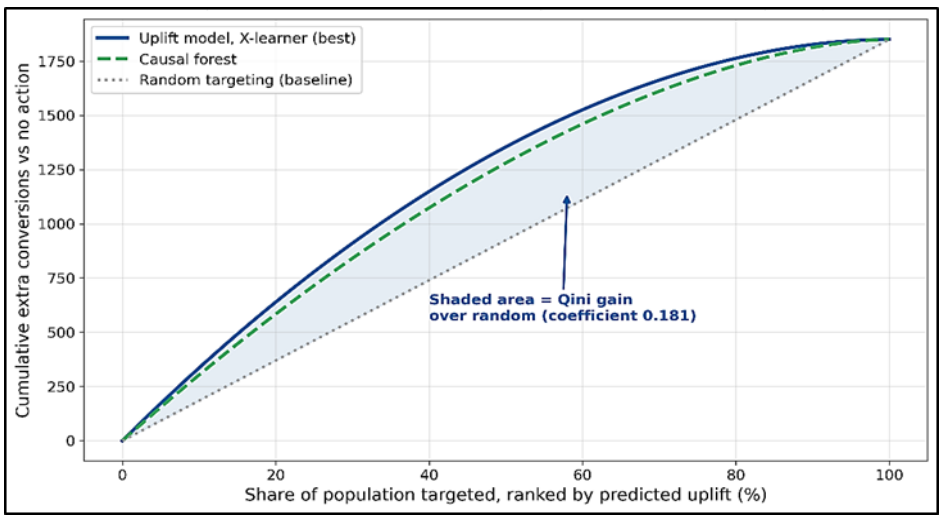


Fig. 4: Qini curve comparing uplift-ranked targeting with random targeting on a synthetic campaign. The shaded area above the diagonal is the Qini gain. Takeaway: ranking by predicted uplift earns more incremental conversions per unit of budget than treating people at random.

C. Domains

Causal estimators now appear across sectors. In pricing, demand response to price is a treatment effect, and confounding from inventory or seasonality biases naive elasticity estimates. In marketing, uplift models direct retention budgets toward persuadable customers. In healthcare, conditional effects guide who benefits from a therapy, with strong overlap requirements because clinicians do not assign treatment at random [22]. In public policy, program evaluation rests on quasi-experimental designs such as instruments and discontinuities, and machine learning estimators extend these to rich administrative data while preserving the credibility of the design [13], [23].

VI. ASSUMPTIONS AND LIMITATIONS

Every number in the previous section depends on assumptions that the data cannot verify. The first is unconfoundedness, also called ignorability, which states that treatment is independent of the potential outcomes given the observed covariates [7]. When an unmeasured factor influences both treatment and outcome, adjustment estimates remain biased, and the bias does not shrink with sample size. Sensitivity analysis quantifies how strong such a hidden factor would have to be to overturn a conclusion, and reporting it is now standard practice [24].

The second assumption is overlap, or positivity: every covariate profile must have a non-trivial probability of receiving each treatment. Where overlap fails, no method can compare like with like, and propensity weights explode. Diagnosing the overlap region and restricting inference to it is often wiser than extrapolating a model into covariate space with no support.

Validation poses its own difficulty. Because the counterfactual outcome is never observed, treatment effect estimates cannot be scored against a ground truth the way predictions can. Practitioners rely on synthetic benchmarks with known effects, on calibration of predicted effects within bins, on placebo and refutation tests that should return a null, and on agreement with randomized experiments where these exist [25], [26]. None of these substitutes for a held-out label, so causal evaluation remains an argument from converging evidence rather than a single accuracy figure.

VII. CONCLUSION

Reliable decisions need answers about interventions, and those answers do not follow from predictive accuracy alone. Causal inference provides the framework for stating what is being estimated and when it can be recovered, while machine learning supplies the flexible estimators that make the framework practical on high-dimensional data. Meta-learners, causal forests, double machine learning, and representation-based deep models now estimate average and conditional effects with controlled bias, as the synthetic benchmark illustrated. The remaining frontier is not algorithmic power but assumption management: stating unconfoundedness and overlap clearly, probing their fragility, and validating with the indirect evidence that causal questions allow. Teams that pair these estimators with honest assumption checks can move from models that describe the world to models that inform action.

REFERENCES

[1] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," J. Educ. Psychol., vol. 66, no. 5, pp. 688–701, 1974.
[2] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[3] G. W. Imbens and D. B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. New York, NY, USA: Cambridge Univ. Press, 2015.
[4] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," J. Amer. Stat. Assoc., vol. 113, no. 523, pp. 1228–1242, 2018.
[5] V. Chernozhukov et al., "Double/debiased machine learning for treatment and structural parameters," Econometrics J., vol. 21, no. 1, pp. C1–C68, 2018.
[6] S. Athey and G. W. Imbens, "Machine learning methods that economists should know about," Annu. Rev. Econ., vol. 11, pp. 685–725, 2019.
[7] G. W. Imbens and D. B. Rubin, "Rubin causal model," in The New Palgrave Dictionary of Economics. London, U.K.: Palgrave Macmillan, 2008.
[8] P. W. Holland, "Statistics and causal inference," J. Amer. Stat. Assoc., vol. 81, no. 396, pp. 945–960, 1986.
[9] S. R. Kunzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," Proc. Nat. Acad. Sci. U.S.A., vol. 116, no. 10, pp. 4156–4165, 2019.
[10] J. Pearl, "Causal diagrams for empirical research," Biometrika, vol. 82, no. 4, pp. 669–688, 1995.
[11] J. Pearl, "Comment: Understanding Simpson's paradox," Amer. Stat., vol. 68, no. 1, pp. 8–13, 2014.
[12] T. J. VanderWeele and I. Shpitser, "A new criterion for confounder selection," Biometrics, vol. 67, no. 4, pp. 1406–1413, 2011.
[13] J. D. Angrist, G. W. Imbens, and D. B. Rubin, "Identification of causal effects using instrumental variables," J. Amer. Stat. Assoc., vol. 91, no. 434, pp. 444–455, 1996.

[14] G. W. Imbens and J. D. Angrist, "Identification and estimation of local average treatment effects," Econometrica, vol. 62, no. 2, pp. 467–475, 1994.
[15] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," Ann. Stat., vol. 47, no. 2, pp. 1148–1178, 2019.
[16] H. Bang and J. M. Robins, "Doubly robust estimation in missing data and causal inference models," Biometrics, vol. 61, no. 4, pp. 962–973, 2005.
[17] E. H. Kennedy, "Towards optimal doubly robust estimation of heterogeneous causal effects," Electron. J. Stat., vol. 17, no. 2, pp. 3008–3049, 2023.
[18] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in Proc. 34th Int. Conf. Mach. Learn. (ICML), 2017, pp. 3076–3085.
[19] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 6446–6456.
[20] N. J. Radcliffe and P. D. Surry, "Real-world uplift modelling with significance-based uplift trees," Stochastic Solutions White Paper, 2011.
[21] P. Gutierrez and J.-Y. Gerardy, "Causal inference and uplift modelling: A review of the literature," in Proc. Mach. Learn. Data Sci. Investment Manage., PMLR, vol. 67, pp. 1–13, 2017.
[22] N. Kallus, "Recursive partitioning for personalization using observational data," in Proc. 34th Int. Conf. Mach. Learn. (ICML), 2017, pp. 1789–1798
[23] S. Athey and G. W. Imbens, "The state of applied econometrics: Causality and policy evaluation," J. Econ. Perspect., vol. 31, no. 2, pp. 3–32, 2017.
[24] P. R. Rosenbaum and D. B. Rubin, "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," J. Roy. Stat. Soc., Ser. B, vol. 45, no. 2, pp. 212–218, 1983.
[25] A. Sharma and E. Kiciman, "DoWhy: An end-to-end library for causal inference," arXiv:2011.04216, 2020.
[26] C. Schuler, M. Baiocchi, R. Tibshirani, and N. Shah, "A comparison of methods for model selection when estimating individual treatment effects," arXiv:1804.05146, 2018.