Sensitivity Analysis for Causal ML: A Use Case at Booking.com

Philipp Bach[1], Victor Chernozhukov[2], Carlos Cinelli[3], Lin Jia[4], Sven Klaassen[1,5], Nils Skotara[4], and Martin Spindler[1,5]

[1] University of Hamburg Business School, Moorweidenstr. 18, 20148 Hamburg, Germany philipp.bach@uni-hamburg.de
[2] Department of Economics and Center for Statistics and Data Science, MIT, Cambridge, MA 02142 USA vchern@mit.edu
[3] Department of Statistics, University of Washington, Seattle, WA 98195-4322 cinelli@uw.edu
[4] Booking.com, Oosterdokskade 163, 1011 DL, Netherlands, {lin.jia},{nils.skotara}@booking.com
[5] Economic AI, Nürnberger Str. 262 A, 93059 Regensburg, Germany {klaassen},{spindler}@economicai.com

Abstract. Causal Machine Learning has emerged as a powerful tool for flexibly estimating causal effects from observational data in both industry and academia. However, causal inference from observational data relies on untestable assumptions about the data-generating process, such as the absence of unobserved confounders. When these assumptions are violated, causal effect estimates may become biased, undermining the validity of research findings. In these contexts, sensitivity analysis plays a crucial role, by enabling data scientists to assess the robustness of their findings to plausible violations of unconfoundedness. This paper introduces sensitivity analysis and demonstrates its practical relevance through a (simulated) data example based on a use case at Booking.com. We focus our presentation on a recently proposed method by [9], which derives general non-parametric bounds on biases due to omitted variables, and is fully compatible with (though not limited to) modern inferential tools of Causal Machine Learning. By presenting this use case, we aim to raise awareness of sensitivity analysis and highlight its importance in real-world scenarios.

Keywords: Causal Machine Learning · Sensitivity Analysis · Unobserved Confounding.

1 Introduction

Randomized controlled trials (RCTs) (also commonly known in industry as “A/B tests,” when the treatment consists of two categories) are widely regarded as the gold standard for estimating causal effects. In an RCT, participants are randomly assigned to either a control group (A) or a treatment group (B), thus preventing individuals from self-selecting into a treatment status. This randomization allows for clear attribution of any differences in outcomes between groups directly to the treatment itself, rather than to other confounding factors. However, due to constraints such as high costs, ethical concerns, or practical limitations, conducting such experiments is often infeasible. In these situations, data analysts may then turn to observational data to infer causal effects.

Unlike experimental studies, observational studies are based on historically collected data, such as past transactions, with no experimental control over the treatment assignment process. In these studies, the treatment and outcome may be associated even without a true causal relationship. This association can occur because individuals' treatment choices and outcomes may both be influenced by confounding factors. To draw valid causal inferences, it is essential to account for these factors effectively. One widely used approach to address these concerns is to adjust for observed covariates, aiming to mitigate confounding biases. This can be done flexibly and efficiently with modern tools of Causal Machine Learning [10]. The validity of this approach, however, still relies on the assumption of unconfoundedness, also known as "selection-on-observables" [1]. This assumption posits that, conditional on the observed covariates, the treatment can be considered *as good as randomly assigned*. Importantly, it assumes no unobserved confounding remains. What if this is not true? How biased would our estimates be? To illustrate these concepts, we will explore a practical simulated example inspired by a real use case at Booking.com.[6] Booking.com is one of the world's leading digital travel platforms and offers not only accommodations, but also other products such as flights, rental cars, and other travel-related services. In this application, we are concerned with estimating the causal effect of cross-selling products on customer relationship. In general, cross-selling, i.e., selling additional products to customers that purchase one product, is an important marketing strategy to increase revenue, customer satisfaction and customer loyalty in the long term. In our analysis, the treatment of interest is whether a customer books an ancillary product in addition to their accommodation booking. The outcome measures the number of follow-up accommodation bookings made within six months after the initial booking. Note an A/B test, which could provide direct experimental evidence of the effect of the treatment on the outcome, is not feasible in this context due to high costs and ethical concerns associated with forcing customers to accept products they may not be interested in.[7]

Given the constraints of not being able to conduct a randomized controlled trial, we rely on observational data, where the treatment assignment stems from customers' choices rather than random allocation. This approach introduces potential confounding factors, as customers' decisions to book ancillary products

[6] We note that the use case is presented in a stylized way. Moreover, due to confidentiality concerns, and for the sake of reproducibility, the empirical analysis is based on a simulated data set. The results of this simulation cannot be used to recover the findings from the actual use case at Booking.com.

[7] Other research designs, such as encouragement designs, could be considered.

are influenced by various factors that may also impact their follow-up bookings. For example, the purpose of a trip — such as whether it is for business or leisure — can significantly affect the likelihood of purchasing additional services like a taxi transfer. Business travelers may be more inclined to book such services compared to leisure travelers. However, these same factors also influence the number of follow-up accommodation bookings. To draw valid causal inferences from observational data, it is thus crucial to account for these confounding variables to isolate the true effect of the treatment. But what if some confounding factors were not measured?

A key concern in this application is whether the observed variables are sufficient to account for all relevant confounding factors, or if there are any unobserved variables that could bias the results. For instance, prior customer loyalty might be a significant factor influencing both the decision to purchase ancillary products and the frequency of follow-up bookings. If this factor is not adequately captured in the data, the assumption of unconfoundedness — where treatment is as good as randomly assigned given observed covariates — might be violated. Sensitivity analysis can then be used to assess how robust our conclusions are to such violations. In some cases, it may suggest that causal inferences are still warranted, even if we did not account perfectly for all confounding factors, such as customer loyalty. Our goal in this paper is to introduce the concepts of causal effect identification, estimation and sensitivity analysis, using the flexible tools of Causal Machine Learning, through this practical example.

The rest of the paper is organized as follows. Section 2, briefly reviews identification via covariate adjustment, and estimation using Debiased Machine Learning (DML) [8]. Section 3, motivates sensitivity analysis, and reviews the recent proposal of [9] for sensitivity analysis in Causal Machine Learning. Section 4 demonstrates this approach in the use case from Booking.com. We complement our analysis with a reproducible notebook, and describe the implementation using the `DoubleML` Python library in the Supplementary Material.

2 Observational Causal Inference

2.1 Potential outcomes and causal parameters

Causal inference starts by *defining* the causal effects of interest, which measure how experimental manipulations of a treatment D would influence an outcome Y. These effects are often formulated in terms of *potential outcomes*. For instance, consider a binary treatment variable, where $D = 1$ denotes treatment, and $D = 0$ denotes the control condition. Here, we write $Y(1)$ to denote the potential outcome of an individual if, perhaps contrary to fact, they were assigned to the treatment condition; analogously, $Y(0)$ denotes the potential outcome of an individual if, perhaps contrary to fact, they were assigned to the control condition. The Average Treatment Effect (ATE) measures the average difference in expected outcomes if the treatment were applied to the entire population compared to this average if the treatment were instead withheld:

$$\theta = E[Y(1)] - E[Y(0)].$$

Similarly, the Average Treatment effect on the Treated (ATT) measures the average treatment effect specifically for those individuals who actually receive the treatment:

$$\theta_0 = E[Y(1) \mid D = 1] - E[Y(0) \mid D = 1].$$

Here we will focus on the ATT, as it is the key causal parameter for our business use case at Booking.com. The ATT measures the average impact on follow-up bookings that results from booking an ancillary product, specifically among those who have actually made such a booking. If the treatment effect varies across individuals and individuals self-select into the treatment, the ATT may differ from the ATE.

Note, however, that we do not have any data on the *potential outcomes* under different experimental conditions. We have only observational data on the actual outcome Y, treatment D and observed covariates X. In order to estimate the ATT, we need assumptions that connect the observational data with the counterfactual parameter we are interested in. The task of deciding whether a set of assumptions is sufficient to recover the causal parameter of interest is often referred to as an *identification problem.*

2.2 Identification under unconfoundedness

The ATT can be *identified* from the distribution of observed data $W_s = (Y, D, X)$ under the following assumptions:

1. *Unconfoundedness*: $\{Y(0), Y(1)\} \perp\!\!\!\perp D|X$,
2. *Overlap*: $0 < P(D = 1|X) < 1$,
3. *Consistency*: $Y = Y(D)$.

Unconfoundedness states that given observed covariates X, the treatment is independent of potential outcomes, implying no unobserved selection mechanisms. Overlap ensures a non-zero probability of receiving treatment for all covariate values. Consistency connects observed outcomes with potential outcomes, and implicitly assumes a well-defined treatment without spillover effects or hidden variations in the treatment.[8] The derivation goes as follows:

$$\begin{aligned}
\theta_0 &= E[Y(1) \mid D = 1] - E[Y(0) \mid D = 1] && (1)\\
&= E[Y(1) \mid D = 1] - E[E[Y(0) \mid D = 1, X]|D = 1] && (2)\\
&= E[Y(1) \mid D = 1] - E[E[Y(0) \mid D = 0, X]|D = 1] && (3)\\
&= E[Y \mid D = 1] - E[E[Y \mid D = 0, X]|D = 1] && (4)
\end{aligned}$$

The first line restates the definition of the ATT. The second line uses the law of total probability (and overlap to avoid division by zero). The third line applies the unconfoundedness assumption, and the fourth line uses the consistency assumption.

[8] This is sometimes also called SUTVA—the Stable Unit Treatment Value Assumption.

We have thus established that, given the stated assumptions, the counterfactual parameter θ_0 can be expressed as a population quantity based solely on observational data, specifically:

$$\theta_0 = E[Y \mid D = 1] - E[E[Y \mid D = 0, X]|D = 1].$$

In practice, however, we do not have access to population quantities. Therefore, we must estimate θ_0 from finite samples.

2.3 Estimation with Causal ML

Various estimation procedures can be used for estimating causal effects from data, including matching, regression, and propensity-score-based methods. Here we focus on causal estimation using machine learning (ML) algorithms. While ML methods can be powerful, their naive application may introduce biases in the estimation procedure that prevent proper uncertainty quantification and statistical inference. To illustrate this point, let us rewrite the ATT as:

$$\theta_0 = E[Y|D = 1] - E[g_s(0, X)|D = 1],$$

where here $g_s(D, X) := E[Y|D, X]$ is the conditional expectation of Y given D and X. The first component of the ATT, namely, $E[Y|D = 1]$, can be easily estimated from data with the sample average of the outcome in the treated group. The second component of the ATT, on the other hand, can be harder to estimate. The function $g_s(D, X)$ can be complex or high-dimensional, motivating the use of ML algorithms. A natural estimator for this quantity would be to directly employ ML estimators for $g_s(D, X)$ and estimate the component $E[g_s(0, X)|D = 1]$, of the ATT by taking the empirical mean of $\hat{g}_s(0, X)$ in the subgroup of treated units. This, however, would lead to severe biases and slow convergence, with estimation errors that are not asymptotically normally distributed [8, 3, 2], precluding standard statistical inference.

Hence, ML-based causal estimation requires certain adjustments to work in practice. A leading framework for such adjustments is the Double/Debiased Machine Learning approach proposed [8], which is based on three key ingredients [3]:[9]

1. Neyman Orthogonal Scores,
2. High-quality Machine Learners,
3. Sample Splitting.

Neyman orthogonality alleviates the regularization bias problem by reducing the sensitivity of the target parameter to estimation errors of nuisance functions such as $g_s(D, X)$. Perhaps counter-intuitively, this often involves rewriting the target parameter as a function of two nuisance parameters instead of one. In the

[9] Here, we briefly sketch the building blocks of DML. For more a more formal introduction, we refer to [8]. An intuitive introduction with multiple examples is available from [3].

ATT example, for instance, it involves estimating the propensity score $m_s(X) := E[D|X]$, and estimating the target parameter with a combination of regression adjustment and inverse probability of treatment weighting. Neyman-orthogonal scores are often available from the literature — several examples, including the ATT, are summarized in [8].

The second key ingredient for DML to work is the quality of the machine learning algorithms used to estimate the nuisance parameters, in terms of convergence rates. In general, they need to have convergence rates faster than $n^{-1/4}$ to guarantee that the higher-order biases vanish. Structural conditions for achieving these learning rates are known for many ML estimators (e.g. the Lasso learner under sparsity). In practice this also demands the careful choice of learners and hyperparameters [4].

The third ingredient, sample splitting, requires that the ML estimation of the nuisance parameters are fitted on a partition of the data set separate from the data used for calculating the target causal parameter. The role of the train and test samples can be swapped, which is called *cross-fitting* — the explicit DML algorithm is presented in the Supplementary Material. [8] then show that the DML causal estimator $\hat{\theta}_0$, built on these three ingredients, concentrates around the true value θ_0 and is asymptotically normal, thus enabling standard statistical inference.

3 Sensitivity Analysis

3.1 Background

In the previous section, we discussed assumptions that allowed us to estimate the Average Treatment Effect on the Treated from observational data, specifically unconfoundedness, overlap, and consistency. However, in general, observed covariates X might not be sufficient to control for confounding due to the presence of unobserved confounders. Figure 1 illustrates a stylized confounding scenario with both observed confounding X and unobserved confounding U, which prevents an unbiased estimation of the causal parameter of interest.

When the unconfoundedness assumption is not fully met, sensitivity analysis provides a way to understand the potential impact of unobserved confounding. The goal is to quantify the influence of unobserved confounders through sensitivity parameters, which measure the strength of confounding in both the treatment assignment mechanism and the outcome equation (illustrated by the red and blue arrows in Figure 1). This usually involves varying sensitivity parameters and assessing how different scenarios impact our causal effect estimates. This can be done through numerical simulations and recalculations of the causal effect estimate under different scenarios, or by positing parameter values into an analytical bias formula.

There is an immense literature on sensitivity analysis spanning the fields of statistics, econometrics, and computer science [16, 29, 27, 18, 28, 23, 6, 20–22, 30, 5, 19, 7, 17, 25, 26, 15, 13, 11, 32, 14]. This body of work includes a variety of approaches, each with different assumptions regarding the strength and nature of

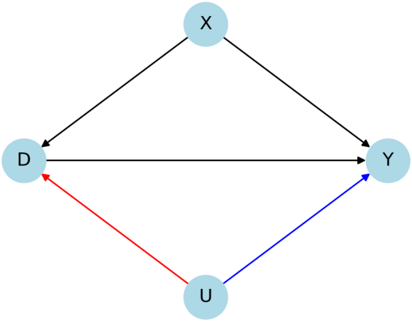

Fig. 1. A stylized DAG illustrating an observed ($D \leftarrow X \rightarrow Y$) and unobserved ($D \leftarrow U \rightarrow Y$) confounding relationship.

unobserved confounding, as well as the specific causal parameters they address. Due to this diversity and complexity, a comprehensive and exhaustive survey of all methods is beyond the scope of this work. Here we focus on the omitted variable bias approach of [9]. This approach is nonparametric, while also permitting (semi-)parametric restrictions (such as partial linearity) when such assumptions are made. It covers a broad range of common causal parameters often investigated in causal inference studies, such as averages of potential outcomes, average treatment effects, average causal derivatives, and policy effects resulting from covariate shifts. Notably, it covers the Average Treatment Effect on the Treated relevant to our use case. This approach also makes it possible to use modern tools of Causal ML for estimation and inference, as described in Section 2.3.

3.2 Omitted Variable Bias in Causal ML

The omitted variable bias approach of [9] starts by positing that unconfoundedness holds given access to the full data $W = \{Y, D, X, U\}$, including the unobserved confounders U, as in Figure 1. Had we had access to this data, we would be able to estimate the target causal parameter of interest (also called the *long* parameter), i.e, the ATT corresponding to an adjustment by X and U,

$$\theta_0 = E[Y|D = 1] - E[E[Y|D = 0, X, U]|D = 1].$$

However, since U is unobserved, the data scientist has access to the observable data only, i.e., $W_s = \{Y, D, X\}$. Similarly, she can only estimate the *short* parameter θ_s, which adjusts for X, but U is *omitted* from the regression equation,

$$\theta_s = E[Y|D = 1] - E[E[Y|D = 0, X]|D = 1].$$

The goal is then to bound the bias due to the omission of U, that is, the difference in the long and short estimands:

$$\text{bias} = \theta_s - \theta_0. \tag{5}$$

[9] show that this bias can be expressed as,

$$\text{bias}^2 = \rho^2 C_Y^2 C_D^2 S^2, \tag{6}$$

where here ρ^2, C_Y^2 and C_D^2 are sensitivity parameters that must be constrained by hypothesis that limit the strength of unobserved confounding, and S^2 is a scaling factor which is estimable from the observed data W_s. In fact, leveraging the Riesz representation theorem, [9] show that the same bias formula holds more generally for any target parameter that can expressed as a linear functional of the conditional expectation function of the outcome. Here we focus on the ATT as it is the relevant parameter for our use case. The general formulation is reviewed in the Supplemental Material.

For the ATT, these sensitivity parameters have the following interpretation. The parameter $\rho^2 \in [0, 1]$ measures the correlation between the confounding errors in the outcome regression $E[Y \mid D, X, U]$ and the treatment regression $E[D \mid X, U]$. For an unobserved variable U to introduce bias in the ATT estimation, it is not enough for U to simply explain variations in both treatment and outcome; the errors in these regressions need to be systematically related. In the absence of additional assumptions about the data-generating process, this parameter is typically set to its upper bound of 1, though less conservative scenarios can also be considered.

The parameters C_Y^2 and C_D^2 quantify how confounding influences the outcome and treatment assignment mechanisms, respectively. They parameterize the blue and red arrows of Figure 1 in terms of the relevant quantities for assessing the omitted variable bias of the ATT. Specifically, $C_Y^2 \in [0, 1]$ is defined as:

$$C_Y^2 := \frac{\mathrm{Var}(E[Y \mid D, X, U]) - \mathrm{Var}(E[Y \mid D, X])}{\mathrm{Var}(Y) - \mathrm{Var}(E[Y \mid D, X])} =: R_Y^2.$$

This quantity, which we also denote by R_Y^2, represents the (nonparametric) partial R^2 of U with respect to Y. In other words, it measures the proportion of the residual variation in the outcome Y that is explained by the unobserved confounder U after accounting for the variation explained by the observed covariates D and X. Notably, R_Y^2 can be set to its upper bound of 1, and this still results in a finite bias.

Finally, the parameter $C_D^2 \in [0, \infty)$ is given by,

$$C_D^2 = \frac{E\left[O(X, U)\right] - E\left[O(X)\right]}{E\left[O(X)\right]},$$

where $O(X, U) := \frac{P(D=1|X,U)}{1-P(D=1|X,U)}$ and $O(X) := \frac{P(D=1|X)}{1-P(D=1|X)}$ represent the odds of receiving treatment, conditional on $\{X, U\}$ and X, respectively. This quantifies the increase in the average odds of receiving treatment due to the presence of the unobserved confounder U. In other words, C_D^2 measures how much the unobserved confounder U improves our ability to predict, on average, whether individuals are in the treated or control group, compared to when we only have access to the observed confounders.

Since C_D^2 is unbounded, it may be useful to re-express it as an R^2-like measure, similar to R_Y^2, for interpretability purposes:

$$C_D^2 = \frac{R_D^2}{1 - R_D^2},$$

where

$$R_D^2 := \frac{E\left[O(X,U)\right] - E\left[O(X)\right]}{E\left[O(X,U)\right]} \in [0,1].$$

The R_D^2 parameter quantifies the proportion of the average odds of receiving treatment explained by the unobserved confounder U, after accounting for what is explained by X. Unlike the other parameters, which can be set to a worst-case value of 1, R_D^2 must always be less than 1 to ensure that the bias remains finite. That is, in order to have informative bounds, we always need to constraint how much confounders affect the treatment assignment.

Estimation and inference All results above hold for population parameters. In practice, we need to estimate the bounds on the target parameter from finite samples. [9] show that estimation and inference for sensitivity analysis can be performed using the same principles of Double/Debiased Machine Learning, as briefly discussed in Section 2.3.

Robustness values The previous bias formula allow us to assess how much bias any confounding scenario, as given by specific values of sensitivity parameters R_Y^2 and R_D^2, would cause. Sometimes, however, researchers may not have a particular scenario in mind. In such cases, users can still report *robustness values* [13, 9], summarising the minimal strength that confounders would need to have in order to revert the research conclusions. Specifically, the robustness value RV_a measures the minimum upper bound on both parameters, $R_Y^2 \leq RV_a$ and $R_D^2 \leq RV_a$, such that the confidence bounds for θ_0 would include a particular value of interest, such as zero, at the significance level a. RV_a thus provide a quick summary of the robustness of an estimated effect—any confounder with $R_Y^2 < RV_a$ and $R_D^2 < RV_a$ is logically incapable of explaining away the observed effect.

Benchmarking Sometimes researchers may not have a good idea about the absolute value of the strength of unobserved confounders, but may instead have a notion of their relative importance when compared to key observed covariates. When that is the case, researchers may use the tool of benchmarking [23, 13, 9], to compute bounds on the strength of unobserved variables if they were as strong or stronger than certain observed covariates.

4 Sensitivity Analysis in a Use Case from Booking.com

We now illustrate the application of the previous sensitivity analysis framework in the use case at Booking.com that was introduced earlier. In this section, we

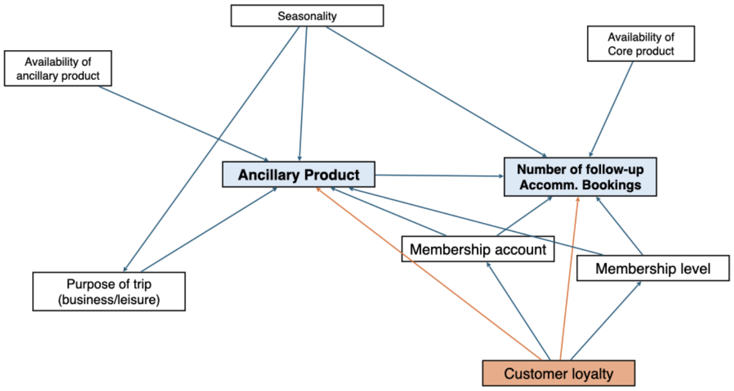
Seasonality
Availability of ancillary product
Availability of Core product
Ancillary Product
Number of follow-up Accomm. Bookings
Membership account
Membership level
Purpose of trip (business/leisure)
Customer loyalty

Fig. 2. Example DAG in the use case.

provide more details on the data set considered and comment on the practical challenges that occurred while applying the omitted variable bias framework of [9]. We complement this analysis with a demonstration notebook that leads through the major steps of the analysis based on the implementation with `DoubleML` for Python [2].[10]

The analysis was based on a pre-analysis plan that defined the sample composition, potential hypotheses, and the definition of the treatment and outcome variable *prior* to the actual estimation step. Customers are defined as visitors of the Booking.com websites and users of the app who have completed an accommodation booking within a time window of six months and have given explicit consent to the usage of their data. The data set is organized in terms of trips, i.e., the purchased ancillary products, were booked for the same trip as the accommodation booking. The original data set is relatively large with a total number of observations > 20 million trips considered. The outcome variable is the number of follow-up bookings of the core product *accommodation* that occurred in a time window of 6 months after the first study period.

The causal analysis was organized according to the workflow displayed in Figure 3. The first step involves a clear and actionable formulation of the causal question, which is derived from business and management considerations. The primary objective of the evaluation was to determine whether the purchase of ancillary products positively impacts customer relationship, as determined by the ATT. A more precise understanding of the causal model and the potential confounding variables that could threaten identification in our use case was achieved by brainstorming directed acyclic graphs (DAGs) representing the various causal relationships of interest. Although identifying a single DAG that perfectly fits

[10] `[Insert url here]`. More information on the implementation is provided in the Supplementary Material.

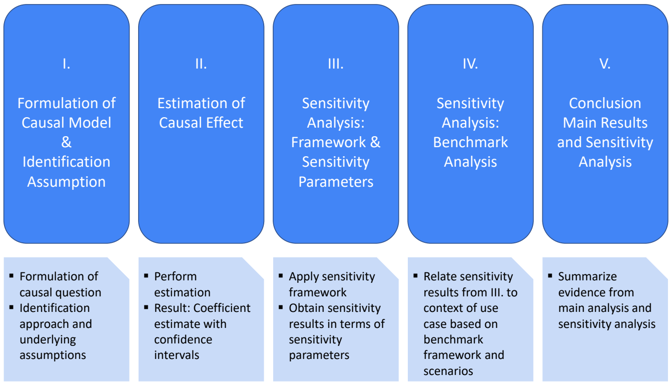
I.
Formulation of Causal Model & Identification Assumption
II.
Estimation of Causal Effect
III.
Sensitivity Analysis: Framework & Sensitivity Parameters
IV.
Sensitivity Analysis: Benchmark Analysis
V.
Conclusion Main Results and Sensitivity Analysis
▪ Formulation of causal question
▪ Identification approach and underlying assumptions
▪ Perform estimation
▪ Result: Coefficient estimate with confidence intervals
▪ Apply sensitivity framework
▪ Obtain sensitivity results in terms of sensitivity parameters
▪ Relate sensitivity results from III. to context of use case based on benchmark framework and scenarios
▪ Summarize evidence from main analysis and sensitivity analysis

Fig. 3. A workflow for causal analysis.

a specific causal problem can be challenging, the process of modeling and discussing these relationships is crucial for formulating the explicit assumptions necessary to identify the causal effect of interest [12]. Figure 2 shows a DAG that maps the main confounders that should be considered for identification via covariate adjustment. Note that other approaches to identification, such as instrumental variables, could also be used, but for simplicity, we do not consider them here. As the DAG shows, the main confounding variables are *Seasonality*, *Purpose of trip*, *Membership account*, *Membership level*, and *Customer Loyalty*, which is not measured, but partially captured by observed covariates. If the arrows in orange do not exist, then adjusting for the observed confounders would be sufficient for identification of the ATT. On the other hand, if these arrows do exist, this means observed variables are not sufficient to account for customer loyalty, and residual biases may still remain.

Assuming unconfoundedness holds, the second step in the workflow involves estimation, which has been performed using Double Machine Learning through the `DoubleML` implementation [2]. For estimating the nuisance functions, `LightGBM` [24] was employed, with parameter tuning based on cross-validation. We obtained an estimate for the ATT of 0.123 with a confidence interval of $[0.107, 0.139]$, indicating a positive, significant, and substantial causal effect for those who purchased an ancillary product. We reiterate that the numerical results presented here do not reflect actual results from the original analysis at Booking.com. The results are based on a simulated data example, which can be replicated using the accompanying notebook.

The previous estimate relies on the assumption of unconfoundedness; specifically it assumes that the orange arrows in the DAG shown in Figure 2 do not exist. The third step of the workflow thus involves performing a sensitivity anal-

Sensitivity Analysis: Summary					
Significance Level		level=0.95			
Sensitivity parameters		R_Y^2=0.110; R_D^2= .003, ρ=1.0			
Bounds with CI					
	CI lower	theta lower	theta	theta upper	CI upper
d	0.070	0.084	0.123	0.162	0.176
Robustness Values					
	H_0	RV (%)	RVa (%)		
d	0.0	5.391	4.816		

Table 1. Results from sensitivity analysis based on preferred benchmarking setting.

ysis with the tools of [9].[11] The results from the sensitivity analysis are presented in Table 1. In the use case, loyalty-based selection into ancillary product purchases are expected to lead to an overestimation of the ATT. Hence, we focus on the lower bound for the ATT and the corresponding confidence bounds to assess whether the effect estimate would be equal or even smaller than zero or become non-significant in a specific confounding scenario. A confounding scenario is defined in terms of fixed values for the sensitivity parameters R_Y^2 and R_D^2. These values can either be derived from domain-specific reasoning or being based on benchmarking exercises. In the absence of a specific scenario, the robustness values provide a concise summary of the robustness of the ATT estimate. For instance, the $RV = 5.391\%$ shows that unobserved confounders that explain less than 5.391% of the residual variation of the outcome and of the odds of treatment, are logically incapable of bringing the point estimate of the ATT to zero. If we consider sampling uncertainty, this number reduces to $RV_a = 4.816\%$ (at the 5% significance level).

Table 1 also shows our preferred confounding scenario, based on benchmarking unobserved confounding against an observed covariate of intermediate strength. In particular, it illustrates how much bias an omitted variable would cause, if it were as strong as "membership variables." These variables were chosen because they are similar in nature to customer loyalty. This scenario results in calibrated values for the sensitivity parameters of $R_Y^2 = 0.110$ and $R_D^2 = 0.003$. In that scenario, the lower bound of the ATT is 0.084 and the bound for the lower confidence limit is 0.070. This is interpreted as evidence in favor of a positive ATT, which would be robust to unobserved confounding similar the benchmarking variable. Note that in this simulated example, the true value of the ATT is 0.070. Thus, in this particular case, the lower limit of the confidence bound provides a better approximation of the true value when compared to a naive estimate that assumes unconfoundedness. In practice, however, the truth is not known. In the actual use case, we also considered stronger, but more implausible, confounding scenarios, benchmarking against all loyalty proxy measures. In the most extreme setting, the ATT was not found to be robust to confounding bias,

[11] We also employed other sensitivity approaches e.g., [31]. A comparison is available in the accompanying notebook.

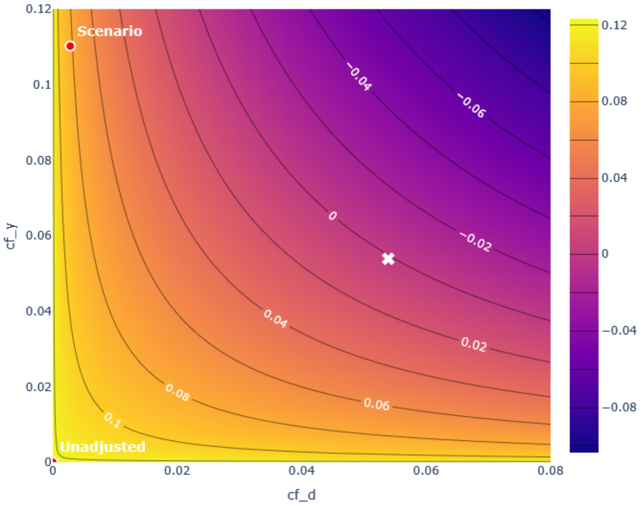

Fig. 4. Contour plot for sensitivity analysis in the cross-selling use case with benchmarking scenario (red point) and the robustness value (white cross).

i.e., the lower bound estimates became negative. An appealing way to visualize multiple, possibly asymmetric confounding scenarios is through a sensitivity contour plot [23, 13, 14]. In Figure 4, a contour line illustrates all combinations of R^2_Y and R^2_D (denoted here as `cf_y` and `cf_d`) that result in the same lower bias limit of the confidence bound for the causal parameter. The conclusions from sensitivity analysis crucially depends on whether the confounding scenarios are realistic and relevant in a specific use case. Data scientists and domain experts evaluated the plausibility of the different scenarios. They concluded that benchmarking against all loyalty measures would correspond to an implausibly strong confounding setting. The results of multiple benchmarking exercises of intermediate strengths led to similar results and, hence, to the conclusion of a rather robust and positive ATT. Finally, integrating sensitivity analysis into the causal analysis workflow has significantly impacted both the analysis and communication of results with business stakeholders. Sensitivity considerations have clarified the fundamental challenges of causal inference with non-experimental data, emphasizing the importance of understanding observational causal analysis and its underlying assumptions. Additionally, sensitivity analysis offered quantitative insights that went beyond intuitive and anecdotal critiques of the unconfoundedness assumption. It provided a clearer understanding of how significant violations of this assumption might be, using domain-specific knowledge relevant to the use case. The sensitivity framework includes practical steps such as calibrating parameters and evaluating benchmark scenarios. By applying sensitivity analysis, data scientists and business stakeholders have gained practical experience that can be applied to future cases.

5 Conclusion and Outlook

In this paper, we introduced causal estimation and sensitivity analysis using machine learning techniques through an applied example that mimics a use case of Booking.com. As ML-based causal analysis becomes increasingly prevalent in both academic research and industry applications, integrating sensitivity analysis is essential. It explicitly addresses the risk of violations of identification assumptions, transparently revealing the threats to the validity of observational studies. The framework developed by [9] offers a powerful formal approach to managing sensitivity concerns, while being fully compatible with modern Causal Machine Learning. By demonstrating this approach through a practical example, we aim to highlight its utility and relevance, encouraging broader adoption of sensitivity analysis in applied research.

References

1. Angrist, J.D., Pischke, J.S.: Mostly harmless econometrics: An empiricist's companion. Princeton university press (2009)
2. Bach, P., Chernozhukov, V., Kurz, M.S., Spindler, M.: DoubleML – An object-oriented implementation of double machine learning in Python. Journal of Machine Learning Research **23**(53), 1–6 (2022), http://jmlr.org/papers/v23/21-0862.html
3. Bach, P., Kurz, M.S., Chernozhukov, V., Spindler, M., Klaassen, S.: DoubleML: An object-oriented implementation of double machine learning in R. Journal of Statistical Software **108**(3), 1–56 (2024). https://doi.org/10.18637/jss.v108.i03, arXiv:https://arxiv.org/abs/2103.096032103.09603 [stat.ML]
4. Bach, P., Schacht, O., Chernozhukov, V., Klaassen, S., Spindler, M.: Hyperparameter tuning for causal inference with double machine learning: A simulation study (2024)
5. Blackwell, M.: A selection bias approach to sensitivity analysis for causal effects. Political Analysis **22**(2), 169–182 (2013)
6. Brumback, B.A., Hernán, M.A., Haneuse, S.J., Robins, J.M.: Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Statistics in medicine **23**(5), 749–767 (2004)
7. Carnegie, N.B., Harada, M., Hill, J.L.: Assessing sensitivity to unmeasured confounding using a simulated potential confounder. Journal of Research on Educational Effectiveness **9**(3), 395–420 (2016)
8. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J.: Double/debiased machine learning for treatment and structural parameters (2018)
9. Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., Syrgkanis, V.: Long story short: Omitted variable bias in causal machine learning (2023)
10. Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., Syrgkanis, V.: Applied causal inference powered by ml and ai (2024)
11. Cinelli, C., Ferwerda, J., Hazlett, C.: sensemakr: Sensitivity analysis tools for ols in r and stata. Available at SSRN 3588978 (2020)
12. Cinelli, C., Forney, A., Pearl, J.: A crash course in good and bad controls. Sociological Methods & Research p. 00491241221099552 (2022)

13. Cinelli, C., Hazlett, C.: Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society Series B: Statistical Methodology **82**(1), 39–67 (2020)
14. Cinelli, C., Hazlett, C.: An omitted variable bias framework for sensitivity analysis of instrumental variables. Available at SSRN 4217915 (2022)
15. Cinelli, C., Kumor, D., Chen, B., Pearl, J., Bareinboim, E.: Sensitivity analysis of linear structural causal models. International Conference on Machine Learning (2019)
16. Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B., Wynder, E.L.: Smoking and lung cancer: recent evidence and a discussion of some questions. journal of National Cancer Institute (23), 173–203 (1959)
17. Dorie, V., Harada, M., Carnegie, N.B., Hill, J.: A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. Statistics in medicine **35**(20), 3453–3470 (2016)
18. Frank, K.A.: Impact of a confounding variable on a regression coefficient. Sociological Methods & Research **29**(2), 147–194 (2000)
19. Frank, K.A., Maroulis, S.J., Duong, M.Q., Kelcey, B.M.: What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. Educational Evaluation and Policy Analysis **35**(4), 437–460 (2013)
20. Frank, K.A., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A., McCrory, R.: Does nbpts certification affect the number of colleagues a teacher helps with instructional matters? Educational Evaluation and Policy Analysis **30**(1), 3–30 (2008)
21. Hosman, C.A., Hansen, B.B., Holland, P.W.: The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. The Annals of Applied Statistics pp. 849–870 (2010)
22. Imai, K., Keele, L., Yamamoto, T., et al.: Identification, inference and sensitivity analysis for causal mediation effects. Statistical science **25**(1), 51–71 (2010)
23. Imbens, G.W.: Sensitivity to exogeneity assumptions in program evaluation. The American Economic Review **93**(2), 126–132 (2003)
24. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30** (2017)
25. Middleton, J.A., Scott, M.A., Diakow, R., Hill, J.L.: Bias amplification and bias unmasking. Political Analysis **24**(3), 307–323 (2016)
26. Oster, E.: Unobservable selection and coefficient stability: Theory and evidence. Journal of Business & Economic Statistics pp. 1–18 (2017)
27. Robins, J.M.: Association, causation, and marginal structural models. Synthese **121**(1), 151–179 (1999)
28. Rosenbaum, P.R.: Observational studies. In: Observational studies, pp. 1–17. Springer (2002)
29. Rosenbaum, P.R., Rubin, D.B.: Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society. Series B (Methodological) pp. 212–218 (1983)
30. Vanderweele, T.J., Arah, O.A.: Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology (Cambridge, Mass.) **22**(1), 42–52 (Jan 2011). https://doi.org/10.1097/ede.0b013e3181f74493, http://dx.doi.org/10.1097/ede.0b013e3181f74493

31. VanderWeele, T.J., Arah, O.A.: Unmeasured confounding for general outcomes, treatments, and confounders: bias formulas for sensitivity analysis. Epidemiology (Cambridge, Mass.) **22**(1), 42 (2011)
32. Zhang, C., Cinelli, C., Chen, B., Pearl, J.: Exploiting equality constraints in causal inference. In: International Conference on Artificial Intelligence and Statistics. pp. 1630–1638. PMLR (2021)

Supplementary Material to *Sensitivity Analysis for Causal ML: A Use Case at Booking.com*

1 DML Algorithm

The DML algorithm proceeds as follows. Consider a Neyman orthogonal score $\psi(\theta, W; \eta)$, where $\eta = (g, \alpha)$. Then, given a random sample $(W_i)_{i=1}^N$ of data vectors, (1) randomly partition it into folds $(I_\ell)_{\ell=1}^L$ of approximately equal size. Denote by I_ℓ^c the complement of I_ℓ. (2) For each ℓ, estimate $\widehat{\eta}_\ell = (\widehat{g}_\ell, \widehat{m}_\ell)$ from observations in I_ℓ^c. (3) Estimate θ_0 as a root of:

$$N^{-1} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \psi(\theta, W_i; \widehat{\eta}_\ell) = 0.$$

2 General Sensitivity Analysis

Overall, we are interested in the causal parameter θ_0 that can be identified as a linear functional of the long regression function $g_0 := E[Y|D, X, U]$:

$$\theta_0 = E[m(W, g_0)],$$

where m is a formula that is affine in g_0, W denotes the full data vector, $W = (Y, D, X, U)$.[1]

The key idea of the formulation in [1] is to express the long parameter θ_0 as the inner product of the long regression, and weights α_0,

$$\theta_0 = E[m(W, g_0)] = E[g_0(W)\alpha_0(W)],$$

where $\alpha_0(W)$ is called the Riesz representer of θ_0.

The Riesz representer plays a crucial role for causal and debiased ML as it implements a correction for the regularization bias that arises from using ML learners, see [3]. In many cases, it is possible to obtain an analytical expression for the Riesz representer. In cases where an analytical debiased presentation is not available, an automated procedure can be performed [3].

The expression for the long parameter θ_0 is based on the complete data that includes information on the unobserved confounder(s). To compare the value of the feasible short estimand to the long parameter, we also reformulate θ_s in terms of its Riesz representation,

$$\theta_s = E[m(W_s, g_s)] = E[g_s(W_s)\alpha_s(W_s)],$$

[1] Here, we abstract from the technical details. For a complete presentation, we refer to [1].

where the subscript s indicates that the quantities refer to the limited available data in the short model, namely, $g_s := E[Y|D, X]$ and $W_s = (Y, D, X)$.

[1] show that the difference between the long and short parameters is given by the covariance of approximation errors in the regression function and the Riesz representer,

$$\theta_s - \theta_0 = E\big[(g_s - g_0)(\alpha_s - \alpha_0)\big].$$

Furthermore, the squared bias can be further expressed as,

$$|\theta_s - \theta_0|^2 = \rho^2 E\big[(g_0 - g_s)^2\big] E\big[(\alpha_0 - \alpha_s)^2\big],$$

where $\rho^2 \in [0, 1]$, is given by the correlation of errors,

$$\rho^2 := \mathrm{Cor}^2(g_0 - g_s, \alpha_0 - \alpha_s).$$

Intuitively, the bias depends on the additional variation that latent variables create in the long regression and the Riesz representer. This bias formula can be further expressed in terms of sensitivity parameters based on R^2 measures, which are more directly interpretable.

$$\begin{aligned} C_Y^2 &:= \frac{E(g_0 - g_s)^2}{E(Y - g_s)^2} = R^2_{Y - g_s \sim g_0 - g_s} \\ C_D^2 &:= \frac{E\alpha_0^2 - E\alpha_s^2}{E\alpha_s^2} = \frac{1 - R^2_{\alpha_0 \sim \alpha_s}}{R^2_{\alpha_0 \sim \alpha_s}}. \end{aligned}$$

In a nonparametric model, $C_Y^2 = R^2_{Y - g_s \sim g_0 - g_s}$ denotes the (nonparametric) partial R^2 of the confounders with the outcome. A similar interpretation applies to $1 - R^2_{\alpha_0 \sim \alpha_s}$, but with respect to the Riesz representer. The term $1 - R^2_{\alpha_0 \sim \alpha_s}$ measures the share of the residual variation in the long version of the Riesz representer that is explained by the confounding variable(s) U. An appealing feature of these sensitivity parameters is that they are naturally restricted to a range between 0 and 1, which improves the interpretability and applicability of the sensitivity framework.

The interpretation of these sensitivity parameters can be further refined depending on the causal parameter of interest, and whether additional parametric assumptions are made. For example, in the setting of prior work by [4], where the target parameter is a linear projection coefficient, both parameters reduce to traditional linear projection partial R^2 measures, $C_Y^2 = R^2_{Y \sim U|D,X}$ and $1 - R^2_{\alpha_0 \sim \alpha_s} = R^2_{D \sim U|X}$, that is, the partial R^2 of U with the treatment D and the outcome Y. This partial R^2 interpretation is still preserved in a high-dimensional partially linear regression model, cf. [1] and [2].

We now provide other examples below.

3 Example: ATE

In the example of a nonparametric causal model under unconfoundedness, we have that the ATE is defined as

$$\theta_0 = E[Y(1) - Y(0)] = E[\underbrace{g_0(1, X, U) - g_0(0, X, U)}_{:=m(W, g_0)}]. \tag{1}$$

The long regression function is

$$g_0(d, X, U) = E[Y|D = d, X, U],$$

with the short version being

$$g_s(d, X) = E[Y|D = d, X].$$

The long Riesz representer for the ATE is

$$\alpha_0(W) = \frac{D}{m_0(X, U)} - \frac{1 - D}{1 - m_0(X, U)},$$

where $m_0(X, U) := E[D|X, U] = P(D = 1|X, U)$ is the long propensity score. Note the RR corresponds to the well known Horvitz-Thompson weights of inverse probability weighting. The short Riesz representer is given by the same weights, excluding U, that is,

$$\alpha_s(W_s) = \frac{D}{m_s(X)} - \frac{1 - D}{1 - m_s(X)},$$

where $m_s(X) := E[D|X] = P(D = 1|X)$ refers to the propensity score without the unobserved confounding variables U.

The sensitivity parameters for the ATE examples then take the following form

$$C_Y^2 := \frac{\mathrm{Var}(E[Y|D, X, U]) - \mathrm{Var}(E[Y|D, X])}{\mathrm{Var}(Y) - \mathrm{Var}(E[Y|D, X])},$$

which equals the non-parametric partial R^2 of Y with U; and,

$$1 - R^2_{\alpha \sim \alpha_s} = \frac{E[1/\mathrm{Var}(D|X, U)] - E[1/\mathrm{Var}(D|X)]}{E[1/\mathrm{Var}(D|X, U)]},$$

which has an interpretation as the gains in precision in the treatment equation due to U.

4 Example: ATT

Here we provide the Riesz representers for the ATT example discussed in the main text:

$$\alpha_0(W) = \left(\frac{D}{m_0(X, U)} - \frac{1 - D}{1 - m_0(X, U)}\right)\left(\frac{m_0(X, U)}{p}\right),$$

$$\alpha_s(W_s) = \left(\frac{D}{m_s(X)} - \frac{1 - D}{1 - m_s(X)}\right)\left(\frac{m_s(X)}{p}\right),$$

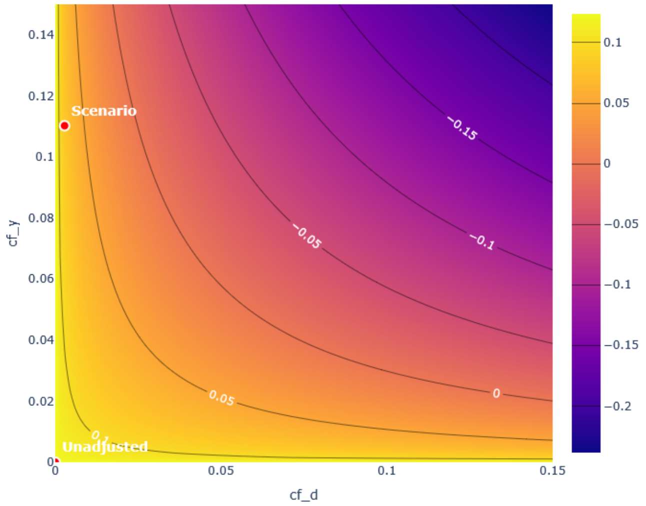

Fig. 1. Contour plot for sensitivity analysis in the cross-selling use case with benchmarking scenario (red point) and the robustness value (white cross) using a calibrated value obtained from benchmarking, $\rho = 0.327$.

where $p := P(D = 1)$. The interpretation of C_Y^2, C_D^2 and $1 - R^2_{\alpha \sim \alpha_s}$ are given in the main text. Note that for simplicity, and to avoid discussing Riesz representation in the main text, the quantity corresponding to $1 - R^2_{\alpha \sim \alpha_s}$ of the general formulation is referred simply as R_D^2.

5 Additional Empirical Results

In the main text, we have not elaborated on the role of the parameter ρ due to space restrictions. The sensitivity parameters C_D^2 and C_Y^2 serve as measures for the variance in the outcome and treatment variable due to an unobserved confounder. The implications of the confounder are not only depending on those quantities, but also to the degree of how much these variations are related to each other. In the case, where U creates variation in D, which, however, is independent from the variation in Y that is induced by U, we would have $\rho = 0$, i.e., such a setting would not alter the estimate of the causal estimate. In empirical examples, the value of ρ can be calibrated, for example through benchmarking. In our analyses, we usually started with the setting of $\rho = 1$ and reduced the value down to the value that was calibrated using benchmarking. Figure 1 illustrates one example with a calibrated value for $\rho = 0.327$, which as been obtained from benchmarking. In line with the general intuition, the results are less conservative.

The contour plot in Figure 4 refers to the plug-in estimate of the lower bound of the ATT. It is also possible to construct similar contour plot for the lower limit of the confidence bound of the ATT.

6 Sensitivity Analysis with the DoubleML library in Python

The Python library DoubleML implements the approach of [1] for various causal models, including the nonparametric treatment effect model and a partially linear regression model. Comprehensive documentation and a detailed user guide are available on the project's webpage https://docs.doubleml.org/stable/index.html. After running a causal analysis in DoubleML, i.e., after instantiation and fitting of a causal model[2], it is possible to perform the sensitivity analysis we discussed here by calling sensitivity_analysis(), which takes fixed values for the sensitivity parameters cf_d and cf_y as input and returns the corresponding bias bounds together with robustness values.[3] The values theta lower and theta upper refer to the point estimates of lower bound and upper bound on the causal parameter (for example the ATT or ATE). The values CI lower and CI upper indicate confidence intervals for the bounds of the causal parameter, thus accounting for sampling uncertainty. The robustness values RV and RVa indicate the minimum upper bound on both sensitivity parameters that would suffice to bring the point estimate (or lower limit of the confidence interval) down to a value of zero. Contour plots can be generated by calling sensitivity_plot(). It is possible to manually add confounding scenarios that are then marked by points in the plot. sensitivity_benchmark() implements the benchmarking procedure by repeating the causal estimation by omitting the candidate benchmark variables. The output provides the values for sensitivity parameters based on comparisons against observed covariates.

[2] See the workflow at https://docs.doubleml.org/stable/workflow/workflow.html

[3] Note here cf_y and cf_d refer to $R^2_Y = C^2_Y$ and $R^2_D = 1 - R^2_{\alpha\sim\alpha_s}$ parameters.

References

1. Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., Syrgkanis, V.: Long story short: Omitted variable bias in causal machine learning (2023)
2. Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., Syrgkanis, V.: Applied causal inference powered by ml and ai (2024)
3. Chernozhukov, V., Newey, W.K., Singh, R.: Automatic debiased machine learning of causal and structural effects. Econometrica **90**(3), 967–1027 (2022)
4. Cinelli, C., Hazlett, C.: Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society Series B: Statistical Methodology **82**(1), 39–67 (2020)