CAUSAL INTERPRETATIONS OF BLACK-BOX MODELS

QINGYUAN ZHAO AND TREVOR HASTIE

Department of Statistics, University of Pennsylvania and Department of Statistics, Stanford University

ABSTRACT. Starting from the observation that Friedman's partial dependence plot has exactly the same formula as Pearl's back-door adjustment, we explore the possibility of extracting causal information from black-box models trained by machine learning algorithms. There are three requirements to make causal interpretations: a model with good predictive performance, some domain knowledge in the form of a causal diagram and suitable visualization tools. We provide several illustrative examples and find some interesting causal relations in these datasets.

1. INTRODUCTION

A central task of statistics is to infer the relationship between "predictor variables", commonly denoted by X, and "response variables", Y. Many if not most of the statistical analyses implicitly hold a *determinism* view regarding this relationship: the input variables X go into one side of a black box and the response variables Y come out from the other side. Pictorially, this process can be described by

A common mathematical interpretation of this picture is

$$Y = f(X, \epsilon), \tag{1}$$

where f is the law of nature and ϵ is some random noise. Having observed data that is likely generated from (1), there are two goals in the data analysis:

Science: Extract information about the law of nature—the function f.
Prediction: Predict what the response variables Y are going to be with the predictor variables X revealed to us.

In an eminent article, Breiman (2001b) contrasts two cultures of statistical analysis that emphasize on different goals. The "data modeling culture" assumes a parametric form for f (e.g. generalized linear model). The parameters are often easy to interpret. They are estimated from the data and then used for science and/or prediction. The

E-mail address: qyzhao@wharton.upenn.edu, hastie@stanford.edu.
Date: October 1, 2018.

"algorithmic modeling culture", more commonly known as machine learning, trains complex models (e.g. random forest, neural nets) that approximates f to maximize predictive accuracy. These black-box models often perform significantly better than the parametric models (in terms of prediction) and have achieved tremendous success in applications across many fields (see e.g. Hastie et al., 2009).

However, the results of the black-box models are notoriously difficult to interpret. The machine learning algorithms usually generate a high-dimensional and highly non-linear function $g(x)$ as an approximation to $f(x)$ with many interactions, making the visualization very difficult. Yet this is only a technical challenge. The real challenge is perhaps a conceptual one. For example, one of the most commonly asked question is the importance of a component of X. Jiang and Owen (2002) notice that there are at least three notions of variable importance:

(1) The first notion is to take the black-box function $g(x)$ at its face value and ask which variable x_j has a big impact on $g(x)$. For example, if $g(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ is a linear model, then β_j can be used to measure the importance of x_j given it is properly normalized. For more general $g(x)$, we may want to obtain a functional analysis of variance (ANOVA). See Jiang and Owen (2002) and Hooker (2007) for methods of this kind.
(2) The second notion is to measure the importance of a variable X_j by its contribution to predictive accuracy. For decision trees, Breiman et al. (1984) use the total decrease of node impurity (at nodes split by X_j) as an importance measure of X_j. This criterion can be easily generalized to additive trees such as boosting (Freund and Schapire, 1996, Friedman et al., 2000) and random forests (Breiman, 2001a). Breiman (2001a) proposes to permute the values of X_j and use the degradation of predictive accuracy as a measure of variable importance.
(3) The third notion is causality. If we are able to make an intervention on X_j (change the value of X_j from a to b with the other variables fixed), how much will the value of Y change?

Among the three notions above, only the third is about the science instead of prediction. To the best of our knowledge, causal interpretation of black-box models has not been studied before, though the reverse direction—using machine learning to aid causal inference—is becoming popular in the literature (van der Laan and Rose, 2011, **?**, Athey and Imbens, 2016, Chernozhukov et al., 2018). This paper aims to fill this vacuum and explain when and how we can make causal interpretations after fitting black-box models.

2. Partial dependence plot

Our development starts with a curious coincidence. One of the most used visualization tools of black-box models is the partial dependence plot (PDP) proposed in Friedman (2001). Given the output $g(x)$ of a machine learning algorithm, the partial dependence of g on a subset of variables $X_{\mathcal{S}}$ is defined as (let $\mathcal{C}$ be the complement set

of $\mathcal{S}$)

$$g_{\mathcal{S}}(x_{\mathcal{S}}) = \mathrm{E}_{X_{\mathcal{C}}}[g(x_{\mathcal{S}}, X_{\mathcal{C}})] = \int g(x_{\mathcal{S}}, x_{\mathcal{C}})\, \mathrm{d}P(x_{\mathcal{C}}). \tag{2}$$

That is, the PDP $g_{\mathcal{S}}$ is the expectation of g over the *marginal* distribution of all variables other than $X_{\mathcal{S}}$. This is different from the conditional expectation $\mathrm{E}[g(X_S, X_C)|X_S = x_s]$, where the expectation is taken over the conditional distribution of X_C given $X_S = x_s$. In practice, PDP is simply estimated by averaging over the training data $\{X_i,\ i = 1, \ldots, n\}$ with fixed $x_{\mathcal{S}}$:

$$\bar{g}_{\mathcal{S}}(x_{\mathcal{S}}) = \frac{1}{n} \sum_{i=1}^{n} g(x_{\mathcal{S}}, X_{i\mathcal{C}}).$$

An appealing property of PDP is that it recovers the corresponding individual components if g is additive. For example, if $g(x) = h_{\mathcal{S}}(x_{\mathcal{S}}) + h_{\mathcal{C}}(x_{\mathcal{C}})$, then the PDP $g_{\mathcal{S}}$ is equal to $h_{\mathcal{S}}(x_{\mathcal{S}})$ up to an additive constant. Furthermore, if g is multiplicative $g(x) = h_{\mathcal{S}}(x_{\mathcal{S}}) \cdot h_{\mathcal{C}}(x_{\mathcal{C}})$, then the PDP $g_{\mathcal{S}}$ is equal to $h_{\mathcal{S}}(x_{\mathcal{S}})$ up to a multiplicative constant. These two properties do not hold for conditional expectation.

Interestingly, the equation (2) that defines PDP is exactly the same as the famous back-door adjustment formula of Pearl (1993). To be more precise, Pearl (1993) shows that if the causal relationship of the variables in (X, Y) can be represented by a graph and $X_{\mathcal{C}}$ satisfies a graphical back-door criterion (to be defined in Section 3.2) with respect to $X_{\mathcal{S}}$ and Y, then the *causal effect* of $X_{\mathcal{S}}$ on Y is identifiable and is given by

$$\mathrm{P}(Y|do(X_{\mathcal{S}} = x_{\mathcal{S}})) = \int \mathrm{P}(Y|X_{\mathcal{S}} = x_{\mathcal{S}}, X_{\mathcal{C}} = x_{\mathcal{C}})\, \mathrm{d}P(x_{\mathcal{C}}). \tag{3}$$

Here $\mathrm{P}(Y|do(X_{\mathcal{S}} = x_{\mathcal{S}}))$ stands for the distribution of Y after we make an intervention on $X_{\mathcal{S}}$ that sets it equal to $x_{\mathcal{S}}$ (Pearl, 2009). We can take expectation on both sides of (3) and obtain

$$\mathrm{E}[Y|do(X_{\mathcal{S}} = x_{\mathcal{S}})] = \int \mathrm{E}[Y|X_{\mathcal{S}} = x_{\mathcal{S}}, X_{\mathcal{C}} = x_{\mathcal{C}}]\, \mathrm{d}P(x_{\mathcal{C}}). \tag{4}$$

Typically, the black-box function g is the expectation of the response variable Y. Therefore the definition of PDP (2) appears to be the same as the back-door adjustment formula (4), if the conditioning set $\mathcal{C}$ is the complement of $\mathcal{S}$.

Readers who are more familiar with the potential-outcome notations may interpret $\mathrm{E}[Y|do(X_{\mathcal{S}} = x_{\mathcal{S}})]$ as $\mathrm{E}[Y(x_{\mathcal{S}})]$, where $Y(x_{\mathcal{S}})$ is the potential outcome that would be realized if treatment $x_{\mathcal{S}}$ is received. When $X_{\mathcal{S}}$ is a single binary variable (0 or 1), the difference $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$ is commonly known as the average treatment effect (ATE) in the literature. We refer the reader to Holland (1986) for some introduction to the Neyman-Rubin potential outcome framework and the Ph.D. thesis of Zhao (2016) for an overview of the different languages of causality.

The coincidence above suggests that PDP is perhaps an unintended attempt to causally interpret black-box models. In the rest of this paper, we shall use several illustrative examples to discuss under what circumstances we can make causal interpretations by PDP and other visualization tools for machine learning algorithms.

3. Causal Model

3.1. **Structural Equation Model.** First of all, we need a causal model to talk about causality. In this paper we will use the non-parametric structural equation model (NPSEM) of Pearl (2009, Chapter 5). In the NPSEM framework, each random variable is represented by a node in a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the node set (in our case $\mathcal{V} = \{X_1, X_2, \ldots, X_p, Y\}$) and $\mathcal{E}$ is the edge set. A NPSEM assumes that the observed variables are generated by a system of nonlinear equations with random noise. In our case, the causal model is

$$Y = f(\text{pa}(Y), \epsilon_Y), \tag{5}$$

$$X_j = f_j(\text{pa}(X_j), \epsilon_j), \tag{6}$$

where $\text{pa}(Y)$ is the parent set of Y in the graph $\mathcal{G}$ and the same for $\text{pa}(X_j)$.

Notice that (5) and (6) are different from regression models in the sense that they are *structural* (the law of nature). To make this difference clear, consider the following hypothetical example:

Example 1. Suppose a student's grade is *determined* by the hours she studied via

$$\text{Grade} = \alpha + \beta \cdot (\text{Hours studied}) + \epsilon. \tag{7}$$

This corresponds to the following causal diagram

If we are given the grades of many students and wish to estimate how many hours they studied, we can invert (7) and run a linear regression:

$$\text{Hours studied} = \alpha' + \beta' \cdot \text{Grade} + \epsilon'. \tag{8}$$

Equation (7) is structural but Equation (8) is not. To see this, (7) means that if a student can study one more hour (either voluntarily or asked by her parents), her grade will increase by β on average. However, we cannot make such interpretation for (8). The linear regression (8) may be useful for the teacher to estimate how many hours a student spent on studying, but that time will not change if the teacher gives the student a few more points since "hours studied" is not an effect of "grade" in this causal model. Equation (8) is not structural because it does not have any predictive power in the interventional setting. For more discussion on the differences between a structural model and a regression model, we refer the reader to Bollen and Pearl (2013).

Notice that it is not necessary to assume a structural equation model to derive the back-door adjustment formula (3). Here we use NPSEM mainly because it is easy to explain and is close to what a black-box model tries to capture.

3.2. **The Back-Door Criterion.** Pearl (1993) shows that the adjustment formula (3) is valid if the variables $X_{\mathcal{C}}$ satisfy the following back-door criterion (with respect to $X_{\mathcal{S}}$ and Y) in the DAG $\mathcal{G}$:

(1) No node in $X_{\mathcal{C}}$ is a descendant of $X_{\mathcal{S}}$; and
(2) $X_{\mathcal{C}}$ blocks every "back-door" path between $X_{\mathcal{S}}$ and Y. (A path is any consecutive sequence of edges, ignoring the direction. A back-door path is a path that contains an arrow into $X_{\mathcal{S}}$. A set of variables block or d-separates a path if the path contains a chain $X_i \to X_m \to X_j$ or a fork $X_i \leftarrow X_m X_j$ such that the middle node X_m is in the set, or the path contains a collider $X_i \to X_m \leftarrow X_j$ such that X_m nor its descendant is in the set.)

More details about the back-door criterion can be found in Pearl (2009, Section 3.3). Heuristically, each back-door path corresponds to a common cause of $X_{\mathcal{S}}$ and Y. To compute the causal effect of $X_{\mathcal{S}}$ on Y from observational data, one needs to adjust for all back-door paths including those with hidden variables (often called unmeasured confounders).

Figure 1 gives two examples where we are interested in the causal effect of X_1 on Y. In the left panel, $X_1 \leftarrow X_3 \leftarrow X_4 \to Y$ (in red color) is a back-door path but $X_1 \to X_2 \to Y$ is not. The set $X_{\mathcal{C}}$ to adjust can be $\{X_3\}$ or $\{X_4\}$. In the right panel $X_1 \leftarrow X_4 \to Y$ and $X_1 \leftarrow X_3 \to X_4 \leftarrow X_5 \to Y$ are back-door paths, but $X_1 \to X_2 \leftarrow Y$ is not. In this case, applying the adjustment formula (3) with $X_{\mathcal{C}} = \{X_4\}$ is not enough because X_4 is a collider in the second back-door path.

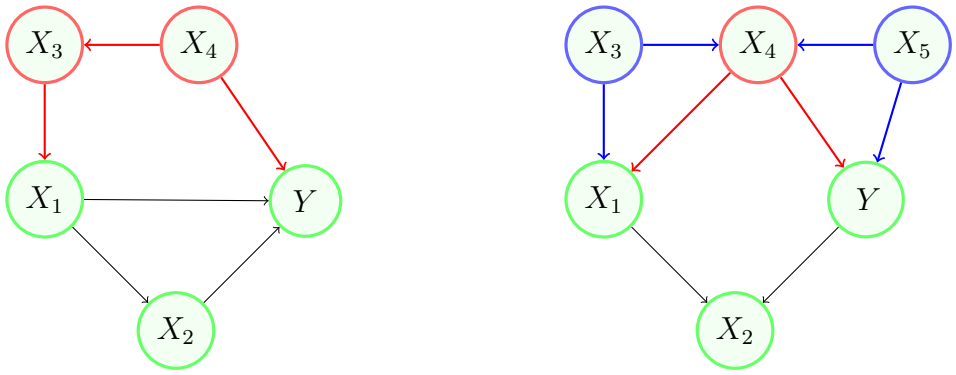


FIGURE 1. Two examples: the red thick edges are back-door paths from X_1 to Y. $\{X_4\}$ blocks all the back-door paths in the left panel but not the right panel (because X_4 is a collider in the path $X_1 \leftarrow X_3 \to X_4 \leftarrow X_5 \to Y$ indicated using the blue color).

Thus the PDP of black-box models estimates the causal effect of $X_{\mathcal{S}}$ on Y, given that the complement set $\mathcal{C}$ satisfies the back-door criterion. This is indeed a fairly strong requirement as no variables in $X_{\mathcal{C}}$ can be a causal descendant of $X_{\mathcal{S}}$. Alternatively if $\mathcal{C}$ does not satisfy the back-door criterion, PDP does not have a clear causal interpretation and domain knowledge is required to select the appropriate set $\mathcal{C}$.

Example 2 (Boston housing data[1]). We next apply PDP for three machine learning algorithms in our first real data example. In an attempt to quantify people's willingness to pay for clean air, Harrison and Rubinfeld (1978) gathered the housing price and other

[1]Taken from `https://archive.ics.uci.edu/ml/datasets/Housing`.

attributes of 506 suburb areas of Boston. The primary variable of interest $X_{\mathcal{S}}$ is the nitrix oxides concentration (NOX, in parts per 10 million) of the areas, and the response variable Y is the median value of owner-occupied homes (MEDV, in \$1000). The other measured variables include the crime rate, proportion of residential/industrial zones, average number of rooms per dwelling, age of the houses, distance to the city center and highways, pupil-teacher ratio, the percentage of blacks and the percentage of lower class.

In order to obtain causal interpretations from the PDP, we shall assume that NOX is not a cause of any other predictor variables.[2] This assumption is quite reasonable as air pollution is most likely a causal descendant of the other variables in the dataset. If we further assume these predictors block all the back-door paths, PDP indeed estimates the causal effect of air quality on housing price.

Three predictive models for the housing price are trained using random forest (Liaw and Wiener, 2002, R package `randomForest`), gradient boosting machine (Ridgeway, 2015, R package `gbm`), and Bayesian additive regression trees (Chipman and McCulloch, 2016, R package `BayesTree`). Figure 2a shows the smoothed scatter plot (top left panel) and the partial dependence plots. The PDPs suggest that the housing price seem to be insensitive to air quality until it reaches certain pollution level around 0.67. The PDP of BART has some abnormal behaviors when NOX is between 0.6 and 0.7. These observations do not support the presumption in the theoretical development in Harrison and Rubinfeld (1978) that the utility of a house is a smooth function of air quality. Whether the drop around 0.67 is actually causal or due to residual confounding requires further investigation.

4. Finer visualization

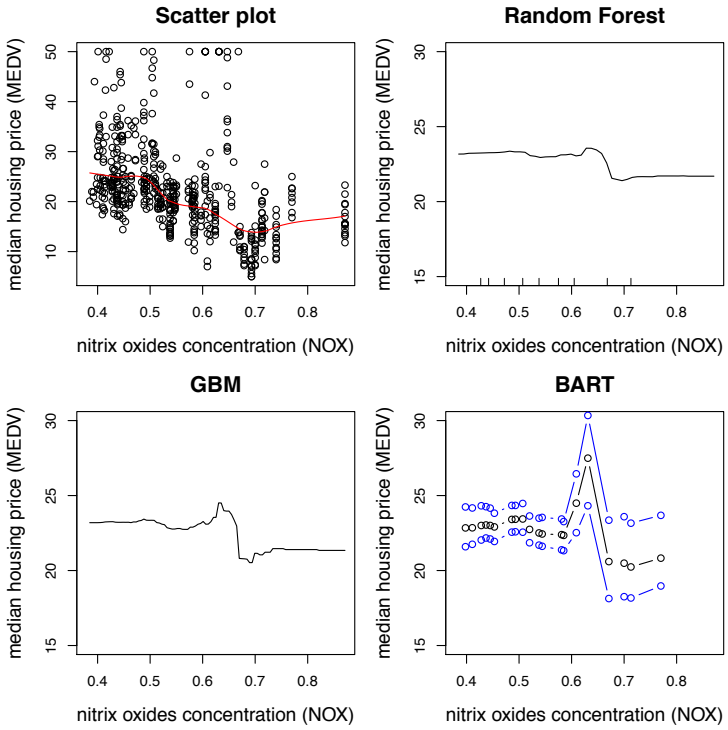
The lesson so far is that we should average the black-box function over the marginal distribution of some appropriate variables $X_{\mathcal{C}}$. A natural question is: if the causal diagram is unavailable and hence the confounder set $\mathcal{C}$ is hard to determine, can we still peek into the black box and give some causal interpretations?

4.1. **Individual Curves.**

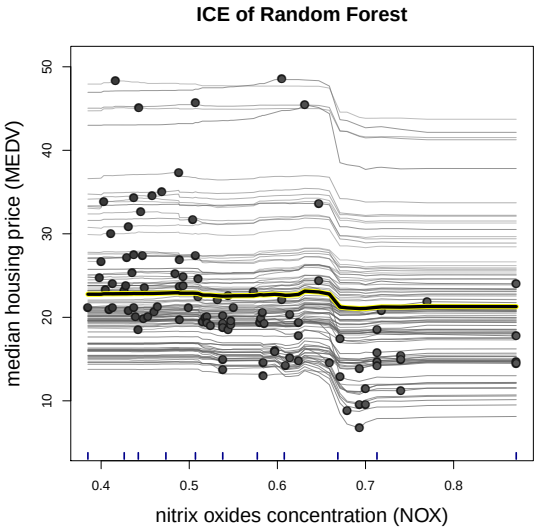
The Individual Conditional Expectation (ICE) of Goldstein et al. (2015) is an extension to PDP and can help us to extract more information about the nature f. Instead of averaging the black-box function $g(x)$ over the marginal distribution of $X_{\mathcal{C}}$, ICE plots the curves $g(x_{\mathcal{S}}, X_{i\mathcal{C}})$ for each $i = 1, \ldots, n$, so PDP is simply the average of all the individual curves. ICE is first introduced to discover interaction between the predictor variables and visually test if the function g is additive (i.e. $g(x) = g_{\mathcal{S}}(x_{\mathcal{S}}) + g_{\mathcal{C}}(x_{\mathcal{C}})$).

Example 3 (Boston housing data, continued). Figure 2b shows the ICE of the black-box model trained by random forest for the Boston housing data. All the individual curves drop sharply around NOX = 0.67 and are quite similar throughout the entire

[2]This statement, together with all other structural assumptions in the real data examples of this paper, are only based on the authors' subjective judgment.



(A) Scatter plot and partial dependence plots using different black-box algorithms. The blue curves in the BART plot are Bayesian credible intervals of the PDP.



(B) ICE plot. (the thick curve in the middle is the average of all the individual curves, i.e. the PDP)

FIGURE 2. Boston housing data: impact of the nitrix oxides concentration (NOX) on the median value of owner-occupied homes (MEDV). The PDPs suggest that the housing price could be (causally) insensitive to air quality until it reaches certain pollution level. The ICE plot indicates that the effect of NOX is roughly additive.

region. This indicates that NOX might have (or might be a proxy for another variable that has) an additive and non-smooth causal impact on housing value.

As a remark, the name "individual conditional expectation" given by Goldstein et al. (2015) can be misleading. If the response Y is truly generated by g (i.e. $g = f$), the ICE curve $g(x_{\mathcal{S}}, X_{i\mathcal{C}})$ is the conditional expectation of Y only if none of $X_{\mathcal{C}}$ is a causal descendant of $X_{\mathcal{S}}$ (the first criterion in the back-door condition). There is perhaps some degree of causal consideration when the name "individual conditional expectation" was invented by Goldstein et al. (2015).

4.2. **Mediation analysis.** In many problems, we already know some variables in the complement set $\mathcal{C}$ are causal descendants of $X_{\mathcal{S}}$, so the back-door criterion in Section 3.2 is not satisfied. If this is the case, quite often we are interested in learning how the causal impact of $X_{\mathcal{S}}$ on Y is *mediated* through these descendants. For example, in the left panel of Figure 1, we may be interested in how much X_1 directly impacts Y and how much X_1 indirectly impacts Y through X_2.

Formally, we can define these causal targets through the NPSEM (Pearl, 2014, VanderWeele, 2015). Let $X_{\mathcal{C}}$ be some variables that satisfy the back-door criterion and $X_{\mathcal{M}}$ be the mediation variables. Suppose $X_{\mathcal{M}}$ is determined by the structural equation $X_{\mathcal{M}} = h(X_{\mathcal{S}}, X_{\mathcal{C}}, \epsilon_{\mathcal{M}})$ and Y is determined by $Y = f(X_{\mathcal{S}}, X_{\mathcal{M}}, X_{\mathcal{C}}, \epsilon)$. In this paper, we are interested in comparing the following two quantities ($x_{\mathcal{S}}$ and $x'_{\mathcal{S}}$ are fixed values):

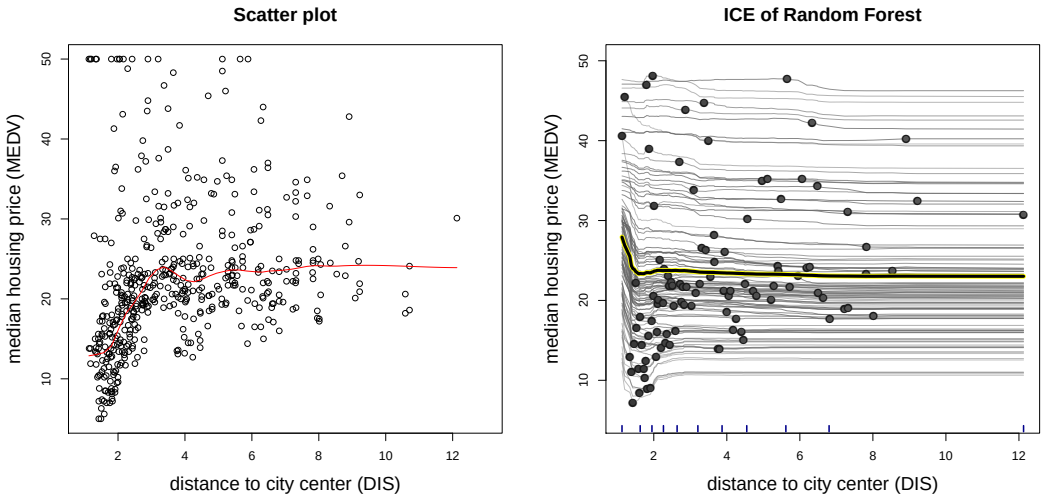
Total effect: $\mathrm{TE} = \mathrm{E}[f(x_{\mathcal{S}}, h(x_{\mathcal{S}}, X_{\mathcal{C}}, \epsilon_{\mathcal{M}}), X_{\mathcal{C}}, \epsilon)] - \mathrm{E}[f(x'_{\mathcal{S}}, h(x'_{\mathcal{S}}, X_{\mathcal{C}}, \epsilon_{\mathcal{M}}), X_{\mathcal{C}}, \epsilon)]$. The expectations are taken over $X_{\mathcal{C}}$, $\epsilon_{\mathcal{M}}$ and ϵ. This is how much $X_{\mathcal{S}}$ causally impacts Y in total.

Controlled direct effect: $\mathrm{CDE}(x_{\mathcal{M}}) = \mathrm{E}[f(x_{\mathcal{S}}, x_{\mathcal{M}}, X_{\mathcal{C}}, \epsilon)] - \mathrm{E}[f(x'_{\mathcal{S}}, x_{\mathcal{M}}, X_{\mathcal{C}}, \epsilon)]$. The expectations are taken over $X_{\mathcal{C}}$ and ϵ. This is how much $X_{\mathcal{S}}$ causally impacts Y when $X_{\mathcal{M}}$ is fixed at $x_{\mathcal{M}}$.

In general, these two quantities can be quite different. When a set $\mathcal{C}$ (not necessarily the complement of $\mathcal{S}$) satisfying the back-door condition is available, we can visualize the total effect by the PDP. For the controlled direct effect, the ICE is more useful since it essentially plots $\mathrm{CDE}(x_{\mathcal{M}})$ at many different levels of $x_{\mathcal{M}}$. When the effect of $X_{\mathcal{S}}$ is additive, i.e. $f(X_{\mathcal{S}}, X_{\mathcal{M}}, X_{\mathcal{C}}, \epsilon) = f_{\mathcal{S}}(X_{\mathcal{S}}) + f_{\mathcal{M},\mathcal{C}}(X_{\mathcal{M}}, X_{\mathcal{C}}, \epsilon)$, the controlled direct effect does not depend on the mediators: $\mathrm{CDE}(x_{\mathcal{M}}) \equiv f_{\mathcal{S}}(x_{\mathcal{S}}) - f_{\mathcal{S}}(x'_{\mathcal{S}})$. The causal interpretation is especially simple in this case.

Example 4 (Boston housing data, continued). Here we consider the causal impact of the weighted distance to five Boston employment centers (DIS) on housing value. Since the geographical location is unlikely a causal descendant of any other variables, the total effect of DIS can be estimated by the conditional distribution of housing price. From the scatter plot in Figure 3a, we can see that the suburban houses are preferred over the houses close to city center. However, this effect is probably indirect (e.g. urban districts may have higher criminal rate, which lowers the housing value). The ICE plot for DIS in Figure 3b shows that the direct effect of DIS has an opposite trend. This suggests that when two districts have the same other attributes, people are indeed willing to pay more for the house closer to city center. However, this effect

is substantial only when the house is very close to the city (DIS < 2), as indicated by Figure 3b.



(A) Scatter plot.

(B) ICE plot. The thick curve in the middle is the average of all the individual curves, i.e. the PDP.

FIGURE 3. Boston housing data: impact of weighted distance to the five Boston employment centers (DIS) on median value of owner-occupied homes (MEDV). The ICE plot shows that longer distance to the city center has a negative causal effect on housing price. This is opposite to the trend in the marginal scatter plot.

5. More examples

Finally, We provide two more examples to illustrate how causal interpretations may be obtained after fitting black-box models.

Example 5 (Auto MPG data[3]). Quinlan (1993) used a dataset of 398 car models from 1970 to 1982 to predict the miles per gallon (MPG) of a car from its number of cylinders, displacement, horsepower, weight, acceleration, model year and origin. Here we investigate the causal impact of acceleration and origin.

First, acceleration (measured by the number of seconds to run 400 meters) is a causal descendant of the other variables, so we can use PDP to visualize its causal effect. The top left panel of Figure 4a shows that acceleration is strongly correlated with MPG. However, this correlation can be largely explained by the other variables. The other three panels of Figure 4a suggest that the causal effect of acceleration on MPG is quite small. However, different black-box algorithms disagree on the trend of this effect. The ICE plot in Figure 4b shows that the effect acceleration perhaps has some interaction

[3]Taken from `https://archive.ics.uci.edu/ml/datasets/Auto+MPG`.

with the other variables (some curves decrease from 15 to 20 while some other curves increase).

Next, origin (US for American, EU for European and JP for Japanese) are causal ancestors of all other variables, so its total effect can be inferred from the box plot in Figure 5a. It is apparent from this plot that Japanese cars have the highest MPG, followed by European cars. However, this does not necessarily mean Japanese manufacturers have the technological advantage of saving fuel. For example, the average displacement of American cars in this dataset is 245.9 cubic centimeters, but this number is only 109.1 and 102.7 for European and Japanese cars. To single out the direct effect of manufacturer origin, we can use the ICE plots of a random forest model, shown in Figure 5b. From these plots, one can see Japanese cars seem to be slightly more fuel-efficient and American cars seem to be slightly less fuel-efficient than European cars even after considering the indirect effects of displacement and other variables.

Example 6 (Online news popularity dataset[4]). Fernandes et al. (2015) gathered 39,797 news article published by Mashable and used 58 predictor variables to predict the number of shares in social networks. For a complete list of the variables, we refer the reader to their dataset page on the UCI machine learning repository. In this example, we study the causal impact of the number of keywords and title sentiment polarity. Since both of them are usually decided near the end of the publication process, we treat all other variables as potential confounders and use the partial dependence plots to estimate the causal effect.

The results are plotted in Figure 6. For the number of keywords, the left panel of Figure 6a shows that it has a positive marginal effect on the number of shares. The PDP in the right panel shows that the actual causal effect might be much smaller and only occur when the number of keywords is less than 4.

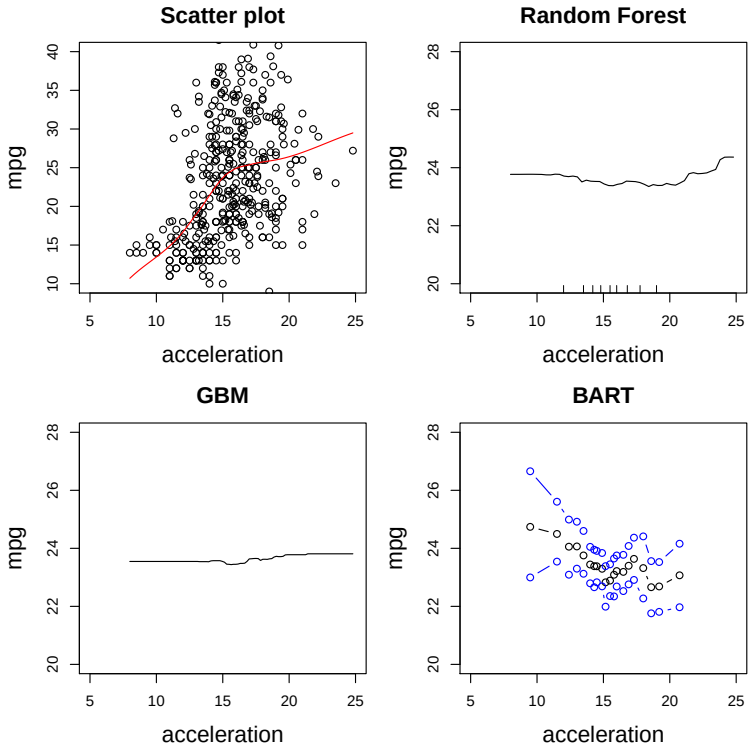
For the title sentiment polarity, both the LOESS plot of conditional expectation and the PDP suggest that articles with more extreme titles get more shares, although the inflection points are different. Interestingly, sentimentally positive titles attract more reshares than negative titles on average. The PDP shows that the causal effect of title sentiment polarity (no more than 10%) is much smaller than the marginal effect (up to 30%) and the effect seems to be symmetric around 0 (neutral title).

6. Conclusion

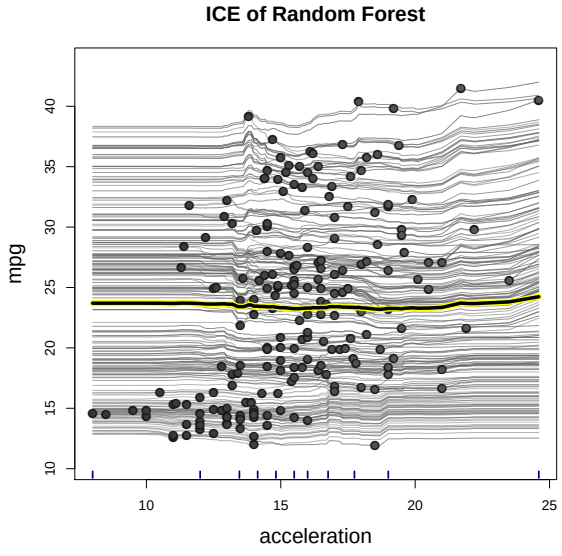
In contrast to the conventional view that machine learning algorithms are just black-box predictive models, we have demonstrated that it is possible to extract causal information from these models using the partial dependence plots (PDP) and the individual conditional expectation (ICE) plots. In summary, we think a successful attempt of causal interpretation requires at least three elements:

(1) A good predictive model, so the estimated black-box function g is (hopefully) close to the law of nature f.

[4]Taken from `https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity`.



(A) Scatter plot and partial dependence plots using different black-box algorithms.



(B) ICE plot. The thick curve in the middle is the average of all the individual curves, i.e. the PDP.

FIGURE 4. Auto MPG data: impact of acceleration (in number of seconds to run 400 meters) on MPG. The PDPs show that the causal effect of acceleration is smaller than what the scatter plot may suggest. The ICE plot shows that there are some interactions between acceleration and other variables.

(2) Some domain knowledge about the causal structure to assure the back-door condition is satisfied.
(3) Visualization tools such as the PDP and its extension ICE.

There are several other directions of research at the crossing of machine learning and causal inference. One relatively well explored direction is to use machine learning to flexibly model nuisance functions in causal inference (van der Laan and Rose, 2011, Chernozhukov et al., 2018). This is important in reducing the model misspecification bias in the semiparametric inference (Dorie et al., 2017). Another useful application of machine learning in causal inference is to aid the discovery of treatment effect heterogeneity (Zhao et al., 2012, Athey and Imbens, 2016). A new and interesting research direction is to use causality to define *fairness* of the machine learning algorithms (Kilbertus et al., 2017, Kusner et al., 2017).

Lastly, we want to emphasize that although PDPs have been shown in the examples to be useful to visualize and possibly make causal interpretations about the black-box models, it should not replace a randomized controlled trial or a carefully designed observational study. Verifying the back-door condition often requires considerable domain knowledge and deliberation, which is usually neglected when collecting data for a predictive task. PDPs can suggest causal hypotheses which should be verified by a more carefully designed study. When a PDP behaves unexpectedly (such as the PDP of BART in Figure 2a), it is important to dig into the data and look for unmeasured confounding. Our hope is this article can encourage more practitioners to peek into their black-box models and discover useful causal relations.

References

Susan C Athey and Guido W Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Kenneth A Bollen and Judea Pearl. Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, pages 301–328. Springer, 2013.

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

Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001b.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Hugh Chipman and Robert McCulloch. *BayesTree: Bayesian Additive Regression Trees*, 2016. URL `https://CRAN.R-project.org/package=BayesTree`. R package version 0.3-1.3.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv preprint arXiv:1707.02641*, 2017.

Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez. A proactive intelligent decision support system for predicting the popularity of online news. In *Portuguese Conference on Artificial Intelligence*, pages 535–546. Springer, 2015.

Yoav Freund and Robert E Schapire. Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference of Machine Learning*, pages 148–156, 1996.

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Elements of Statistical Learning*. Springer, 2009.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16, 2007.

Tao Jiang and Art B Owen. Quasi-regression for visualization and interpretation of black box functions. Technical report, Stanford University, Stanford, 2002.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

Judea Pearl. Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. Interpretation and identification of causal mediation. *Psychological Methods*, 19(4):459, 2014.

J Ross Quinlan. Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 236–243, 1993.

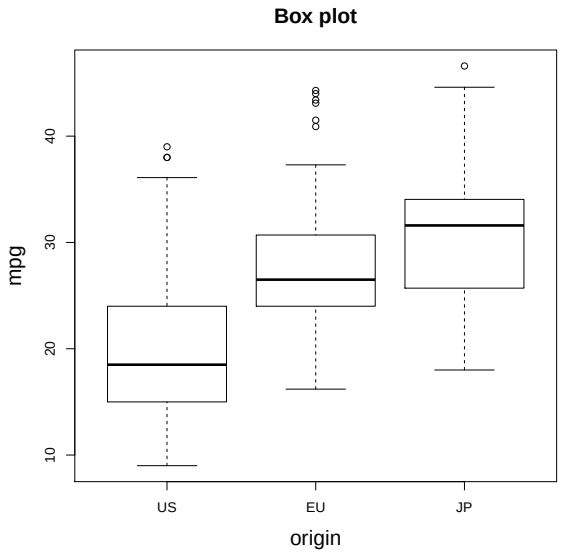
Greg Ridgeway. *gbm: Generalized Boosted Regression Models*, 2015. URL `https://CRAN.R-project.org/package=gbm`. R package version 2.1.1.

Mark J. van der Laan and Sherri Rose. *Targeted Learning*. Springer, 2011.

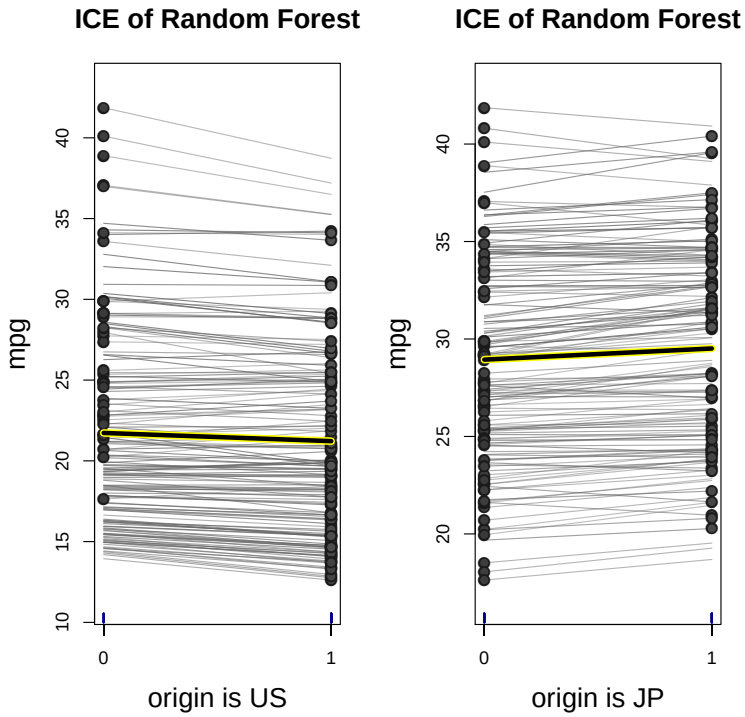
Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

Qingyuan Zhao. *Topics in Causal and High Dimensional Inference*. PhD thesis, Stanford University, 2016.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

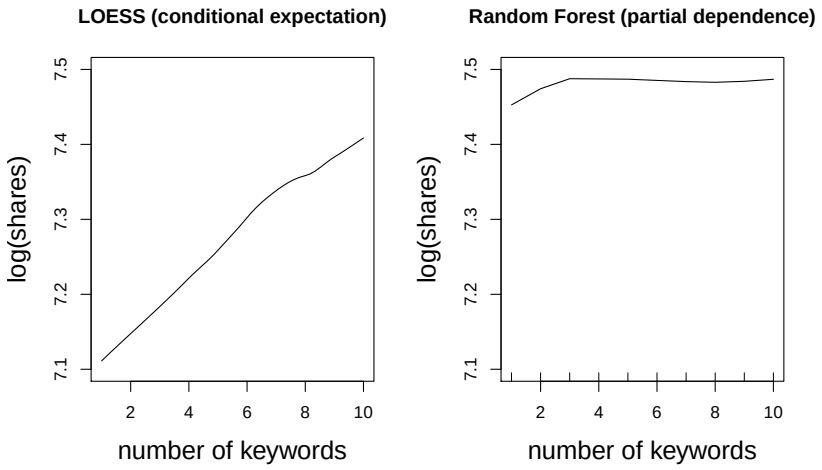


(A) Box plot.

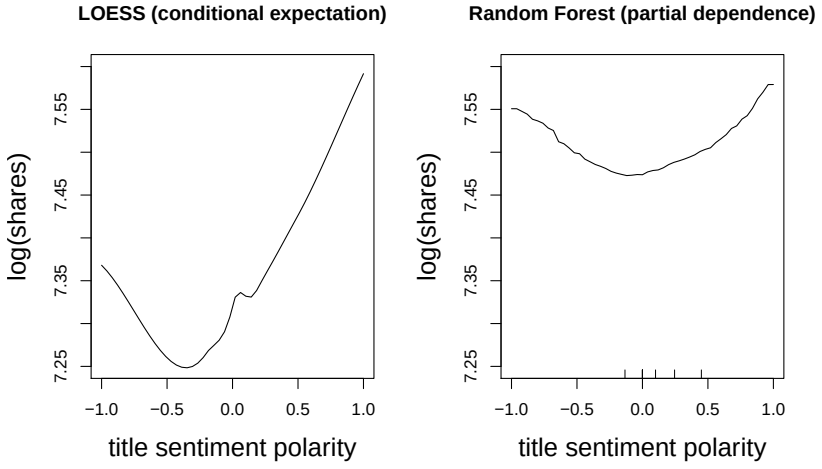


(B) ICE plots. The baseline (level 0) in both plots is origin being EU. For clarity, only 50% of the ICE curves in the left panel are shown.

FIGURE 5. Auto MPG data: impact of origin on MPG. Marginally, Japanese cars have much higher MPG than American cars. This trend is maintained in the ICE plots but the difference is much smaller.



(A) Impact of number of keywords on log of shares. The PDP shows that the actual causal effect might be much smaller than the marginal effect and only occur when the number of keywords is less than 4.



(B) Impact of title sentiment polarity on log of shares. Both plots suggest that extreme titles get more shares. The PDP shows that the causal effect might be much smaller than the marginal effect.

FIGURE 6. Results of online news popularity dataset.