Interpreting Predictive Models through Causality: A Query-Driven Methodology

Mahdi Hadj Ali, Yann Le Biannic, Pierre-Henri Wuillemin

Interpreting Predictive Models through Causality: A Query-Driven Methodology

Mahdi HADJ ALI [1,2] * **Yann LE BIANNIC** [2] **Pierre-Henri WUILLEMIN** [1]
[1] LIP6 (UMR 7606 Sorbonne Université – CNRS), 4 pl. Jussieu 75005 Paris, France.
[2] SAP France, 35 rue d'Alsace, 92300 Levallois-Perret, France.

Abstract

Machine learning algorithms have been widely adopted in recent years due to their efficiency and versatility across many fields. However, the complexity of predictive models has led to a lack of interpretability in automatic decision-making. Recent works have improved general interpretability by estimating the contributions of input features to the prediction of a pre-trained model. Despite these advancements, practitioners still seek to gain causal insights into the underlying data-generating mechanisms. To this end, some works have attempted to integrate causal knowledge into interpretability, as non-causal techniques can lead to paradoxical explanations. These efforts have provided answers to various queries, but relying on a single pre-trained model may result in quantification problems. In this paper, we argue that each causal query requires its own reasoning; thus, a single predictive model is not suited for all questions. Instead, we propose a new framework that prioritizes the query of interest and then derives a query-driven methodology accordingly to the structure of the causal model. It results in a tailored predictive model adapted to the query and an adapted interpretability technique. Specifically, it provides a numerical estimate of causal effects, which allows for accurate answers to explanatory questions when the causal structure is known.

1 Introduction

Recent Machine Learning (ML) methods are increasingly sophisticated and generally improve the accuracy of the models constructed but at the expense of greater difficulty of interpretation. Moreover, the interpretability of these models is a sensitive issue in many fields (Burkart and Huber 2021). Indeed, using models in the context of automatic decision-making requires detailed knowledge of their behavior in order to be able to justify the decision; for instance, in the medical domain of automatic prescription, in the legal domain, or in a legal context (Rieg et al. 2020). Practitioners are often interested in causal insights into the underlying data-generating mechanisms, which Machine Learning methods do not generally provide. Common causal questions include the identification of causes and effects, predicting the effects of interventions, and answering counterfactual questions. If we assume that the underlying causal model of the data generation process can be represented as a causal Bayesian network (i.e. a Bayesian network where orientations have a causal interpretation), the ideal solution is to utilize the causal framework and specialized tools such as do-calculus to answer those queries. However, obtaining the complete causal model can be challenging due to the multiplicity of parents for the target or the impossibility of querying latent variables. Hence, we may have to rely on assumptions only about its causal structure, and on predictive models.

Various works deal with quantifying causal effects (direct and/or indirect) from a predictive model, presuming knowledge of the causal structure, which depicts the connections among features (Heskes et al. 2020; Wang, Wiens, and Lundberg 2021). These studies follow the same pattern as the Explainable AI (XAI) field, i.e. starting with a predictive model, typically trained from all known variables, and then trying to quantify the contribution of each variable. The objective of this paper is to show the benefits of an alternative approach where the predictive model is no longer given but is designed to answer a specific causal query. As in previous works, we assume prior knowledge about the causal structure, but we propose using it before building, training, and analyzing query-driven predictive models from observational data.

The first section of this paper presents state-of-the-art XAI techniques and some notions of causality. Next, we describe our approach and the setup that will allow us to compare the different approaches on a synthetic dataset. Section four shows the limitations of using a predefined predictive model typically trained on all input variables. In the last sections, we will study two causal questions that are difficult to address with current approaches but can be correctly treated by our proposal.

2 Predictive models, causal models

2.1 Predictive models and explainability

A common task in supervised learning is to predict the binary class Y of an object from a vector of features $\mathbf{X} = \{X_1, \cdots, X_j, \cdots, X_M\}$. A **machine-learning model** is trained from a database of observations about the class and the features. It is defined as a real-valued function f that takes a vector of features as input and returns an estimate of

*This work is conducted as part of a CIFRE thesis (no2020/1640) supported by SAP France and ANRT.

the probability of the target class: $f(\mathbf{X}) \simeq P(Y = 1|\mathbf{X})$.

Several tools have been developed to explain the predictions made by an ML model. For instance, the *Partial Dependence Plots* (PDP) proposes to examine the effect of the j-th variable by studying the average prediction when this variable is perturbed (Friedman 2001). The *Individual Conditional Expectation Plots* (ICE) are based on the same idea as the PDPs but correspond to the study of the prediction by f from a given example when the j-th variable is modified (Goldstein et al. 2015). Thus, the average of all the ICEs corresponds to the PDP.

Another idea is to estimate an importance score for each variable. (Breiman 2001) proposes to exchange a variable with noise and assess the impact on predictions.

This paper will often refer to Shapley values (Shapley 1953) and their application to XAI. Shapley values are a method to spread credit among $\mathbf{X}$ players in a "coalition game" (von Neumann and Morgenstern 1947). In this framework, a value function v associates a real number $v(S)$ to any coalition $S \subseteq \mathbf{X}$. To transpose this framework to XAI, a parallel is drawn between the prediction by a model and the value function for a game and between the input features $\mathbf{X}$ and the players who collaborate to gain $f(\mathbf{X})$.

Shapley values thus became a way to explain a model and they spread in the area of ML (Strumbelj and Kononenko 2010; Lundberg and Lee 2017). Several variants (Sundararajan and Najmi 2020; Frye, Rowat, and Feige 2020; Heskes et al. 2020; Wang, Wiens, and Lundberg 2021; Kolpaczki, Bengs, and Hüllermeier 2023) have been proposed. Among these, we will mostly make reference to the widespread SHAP values (Lundberg and Lee 2017).

The explanation provided by SHAP values is an excellent basis for understanding the behavior of a predictive model. SHAP values offer a model-agnostic explanation and are based on solid mathematical foundations. However, the problem of explainability often lies more in prescribing than in predicting. Predictive Analytics aims at answering questions such as "What is the likely value of Y if I observed X?" or "What are the weights of the evidence leading to the prediction?". On the other hand, Prescriptive Analytics address questions such as "What intervention should I do to improve Y?" and "When and why should I make such an intervention?". SHAP values quantify the contributions of features to the prediction made by a model and thus fit the needs of Predictive Analytics. However, the weight of evidence can be easily confused with the effect of an intervention. The latter is needed for Prescriptive Analytics.

2.2 Causal models and explainability

One potential solution to prescriptive questions is to turn to the causal framework and tools. From a causal perspective, these questions can be answered by the causal effect of an actionable variable on the target. An actionable variable is a variable that can be acted upon in the "real world" i.e. one can intervene on the variable and thus control its value. The do-calculus (Pearl 2012) is a solution for estimating causal effects.

Janzing et al. (2013) propose to quantify the causal contribution of a binary variable A by its Average Causal Effect (ACE):

$$\mathbb{E}[Y|do(A = 1)] - \mathbb{E}[Y|do(A = 0)]$$

This is similar to the concept of average uplift (Rubin 1974; Gutierrez and Gérardy 2017; Devriendt, Moldovan, and Verbeke 2018), which is appreciated for its simplicity and thus facilitates decision-making.

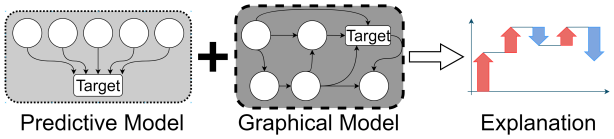

Figure 1: Classic XAI Approach.

2.3 Our proposition to combine causal and predictive models

XAI methods typically aim at explaining the predictions made by a previously trained model. Some methods incorporate causality via a graphical model of the underlying causal relationships between variables (Frye, Rowat, and Feige 2020; Heskes et al. 2020). However, these methods inherit from general XAI the premise that a single pre-trained predictive model is the main source of estimates to answer causal queries about multiple variables, see Figure 1.

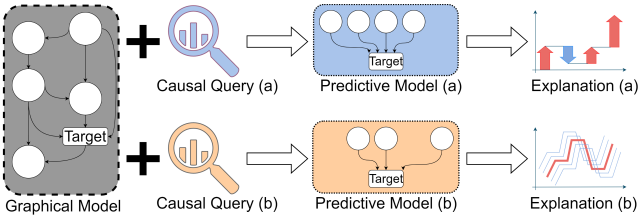

Figure 2: Proposed approach with 2 distinct queries.

In this paper, we propose a new methodology, illustrated in Figure 2, which extends the common framework described above. Our approach consists of several phases. First, we start with a training population, a causal graph, and a specific causal query. Next, we train an ML model tailored to the query and causal context. Finally, we use an interpretability method adapted to the query and context to quantify the desired effect.

A key difference between our proposal and previous methods is that we do not assume a pre-trained model. The main argument is that different causal questions cannot be systematically answered by a single general predictive model. The construction of the model that generates the explanation must also consider the constraints imposed by the causal computation.

3 Experimental protocol

This article examines the feasibility of quantifying multiple causal effects from observational data using standard supervised learning algorithms and interpretable ML techniques.

In practice, the learned models may be biased or distorted. To overcome these problems, we propose a causal Bayesian network as the ground truth reference.

A database is generated from this reference model. This data is used as a learning base for the predictive models we are trying to explain. Using a causal Bayesian network as ground truth allows us to quantify the exact causal effects of the features of interest using analytical methods such as do-calculus (Pearl 2000). Thus, we can examine a classification model's interpretations and assess their consistency with the underlying causal model.

To illustrate our point, we designed a synthetic example with pyAgrum, a library for probabilistic graphical models (Ducamp, Gonzales, and Wuillemin 2020). To facilitate the reasoning, we assigned a semantic to this example: the task of predicting whether the customer will renew his cell phone subscription. The prediction is based on several features:

- *Economy* (noted as E) represents economics conditions, from expansion to contraction,
- the client profile (e.g. residential vs commercial) is represented by the variable *Customer Profile* (noted as C),
- the yearly consumption of the service by the customer is tracked by *Usage* (noted as U). ,
- a one-time offer granted to the customer is illustrated by *Discount* (noted as D),
- the *Loyalty* of the client cannot be directly observed and will be handled as a latent variable (noted as L),
- *Visits* (noted as V) indicates whether the customer has visited the provider website recently,
- finally *Renewal* (noted R) informs about subscription renewal and will be the target for binary classification.

To limit the feature space size and train accurate classification models, most variables are binary except *Usage*, which can take five distinct values. Figure 3 represents the causal Bayesian network used to generate data samples.

Two explanations of interest are the effect of the *Economy* and the *Discount*. The fictitious model has been designed so that granting a discount (*D=1*) has a positive causal effect on renewals for one customer profile (*C=0*) and no causal effect for the other profile (*C=1*):

$$P(R|do(D=1), C=0) > P(R|do(D=0), C=0)$$

$$P(R|do(D=1), C=1) = P(R|do(D=0), C=1)$$

and:

$$P(R|do(D=1)) > P(R|do(D=0))$$

Similarly, *Economy* (*E=1*) has a total negative effect on *Renewal* when *C=0* and no causal effect when *C=1*.

The purpose of the next sections is to show in different contexts how the causal interpretation of classical XAI results can be ambiguous (section 4) and how our proposition can lead to more consistent estimations of causal effects from specific prediction tasks (section 5 and 6). The implementation of the examples is provided as a notebook on GitHub[1].

[1]Figures and models can be found in `https://github.com/anonyme/query-driven-xai`

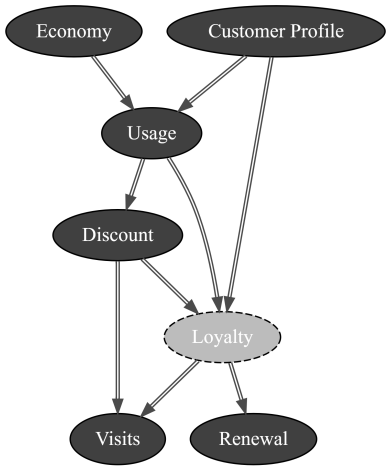

Figure 3: The causal Bayesian network used to generate the dataset. *Loyalty* is considered as a latent variable.

4 Sensitivity to feature selection

In any analysis of observational data, it is well-known that the selection of variables has a significant impact. This section illustrates how this selection without causal analysis can lead to paradoxes (subsection 1) or unnecessary/missing quantification of some parameters (subsection 2).

4.1 Paradoxical insights from general XAI

To train predictive models from populations, we use a well-known ML algorithm, XGBoost (Chen and Guestrin 2016). We train two models on the same dataset but using different sets of features. A first model is trained on all known features (i.e., all variables except the target and the unobserved *Loyalty*), and a second model is trained after dropping *Visits*. For these two models, we use the SHAP library (Lundberg and Lee 2018) to compute the contributions of the features. The results are given in Figure 4 and Figure 5.

Figure 4: Summary Plot from SHAP, explaining a model trained on all variables.

The two plots represent, as documented in the SHAP library, the SHAP values of every feature for every sample. The plot sorts features by the sum of SHAP value magnitudes over all samples and use SHAP values to show the distribution of each feature impacts on the model output. The color represents the feature value (red high, blue low).

A reading of these two plots suggests that granting a discount (*Discount* red dots) contributes negatively to the pre-

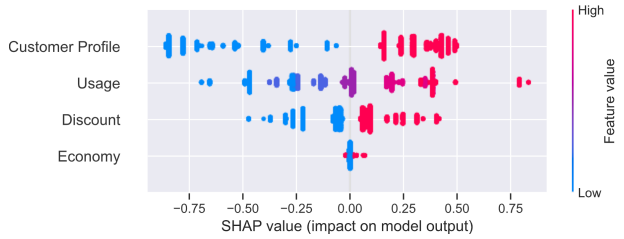

Figure 5: Summary Plot from SHAP, explaining a model trained excluding *Visits*.

dictions in the first model Figure 4, while it has a positive contribution in the second model Figure 5. If contributions were naively interpreted as causal effects on the target, an analyst might draw opposite conclusions from the two models. In this example, we observe that the widespread SHAP interpretation method is sensitive to feature selection: it may provide conflicting insights when applied to different models trained using the same ML algorithm on the same dataset, but on different selections of features.

4.2 Predictive power versus causal effect

Several authors have proposed incorporating causal structure knowledge when interpreting a predictive model. However, quantifying causal effects may require information that cannot be extracted from the model. Indeed, the predictive model may not use a variable that has an indirect causal effect on the target. This situation arises when the variable is independent of the target upon conditioning on other input features.

If we assume that SHAP correctly represents the contributions of the input variables to the predictions (Janzing, Minorics, and Bloebaum 2020), we can observe this situation in our synthetic example. *Economy* has an indirect causal effect on the target: in the data-generating model, its Average Causal Effect (ACE) is about -2.8%. However, the causal effect of *Economy* goes through a mediator (*Usage*) that is an input feature of the predictive model. Thus, *Economy* brings no additional information about the target over *Usage* and can be ignored by the model without any impact on prediction accuracy. Indeed, we observe from SHAP values extracted from both our predictive models that the contribution of *Economy* is close to zero .

On the other hand, a variable may strongly contribute to a predictive model while being neither a direct or indirect cause nor a direct or indirect consequence of the target. In our synthetic example, *Visits* is neither a cause nor a consequence of renewals, but the existence of a latent variable (*Loyalty*) implies that *Visits* is not independent of the target when conditioning on all known variables, and thus *Visits* brings additional information about the target. Indeed, the SHAP plots show that *Visits* is the top predictor of the model that has access to this variable.

5 Quantification of a total causal effect

Let us assume that the objective is to appraise the effect of a discount on subscriber renewal. In this section, we show how to exactly compute this effect under the assumption of the full causal model and then how a query-driven application of XAI tools from observational data allows a reliable approximation of the effect, even in the presence of latent variables.

5.1 Exact solution using do-calculus

Within a probabilistic causal framework, the query for the total causal effect is the quantification of the probability $P(Y|do(X))$. In such a framework, do-calculus provides multiple techniques, such as frontdoor or backdoor adjustments, to compute causal effects (Pearl 2000). In particular, the backdoor adjustment defines a set of variables that should be considered.

Definition *(Backdoor Criterion) — Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y):*

- *(i) if no node in Z is a descendant of X, and*
- *(ii) Z blocks every path between X and Y that contains an arrow into X.*

If a set of variable Z satisfies the backdoor criterion relatively to (X, Y), then the causal effect of X on Y is identifiable and is given by the following adjustment:

Definition *(Backdoor Adjustment) - If Z satisfies the backdoor criterion relative to (X, Y):*

$$P(Y|do(X = x)) = \sum_z P(Y|X = x, Z = z)P(Z = z) \tag{1}$$

Applied to our example (Figure 3), the backdoor adjustment is suitable for quantifying the causal effect of *Discount* on *Renewal* with $\{Usage\}$ as a set satisfying the backdoor criterion.

5.2 Estimates from a sample data

Estimating the causal effect through the backdoor adjustment in Equation 1 only involves the variables Y, X and Z. Equation 1 can be generalized and reformulated using $X_S = \{X\}$, $X_{\bar{S}} = Z$:

$$P(Y|do(x_S)) = \int P(Y|X_S = x_S, X_{\bar{S}} = x_{\bar{S}})dP(x_{\bar{S}})$$

To compute this quantity from observational data, a proper process is to build a probabilistic model f of Y knowing only $\mathbf{X} = X_S \cup X_{\bar{S}}$ and then to rely on a Monte-Carlo integration over the training data where the probability $P(Y|\mathbf{X})$ is estimated by $f(\mathbf{X})$:

$$P(Y|do(x_S)) \simeq \frac{1}{N}\sum_{i=1}^{N} P(Y|X_S = x_S, X_{\bar{S}}^i) \tag{2}$$

$$\simeq \frac{1}{N}\sum_{i=1}^{N} f(x_S, X_{\bar{S}}^i). \tag{3}$$

Zhao and Hastie (2019) already demonstrated the analogy between the backdoor adjustment and the partial dependence plot (PDP).

Given a predictive model $f(\mathbf{X})$, a PDP grants visualization and analysis of the dependence of the predictions on an input feature of interest S (let $\bar{S}$ be its complement). The PDP can be computed as shown in Equation 4.

$$f_S(x_S) = E_{X_{\bar{S}}}[f(x_S, X_{\bar{S}})] = \int f(x_S, x_{\bar{S}}) dP(x_{\bar{S}}) \quad (4)$$

Indeed, the Monte-Carlo integration of Equation 4 over the training data has exactly the same equation as Equation 3.

This development demonstrates that prior causal knowledge guides toward relevant selections of variables for building predictive models so that tools such as PDP acquire a causal meaning.

5.3 Illustration: effect of *Discount* on *Renewal*

By construction, the causal model of the synthetic data generation process grants access to the true causal effect that pyAgrum can compute directly through do-calculus. The calculation involves a backdoor adjustment with $\{Usage\}$ as the minimal set that satisfies the backdoor criterion (see Equation 5). Indeed two sets satisfy the criterion 5.1: $\{Usage\}$ and $\{Usage, Customer\ profile\}$. For the first set Equation 1 becomes:

$$P(R|do(D=d)) = \sum_U P(R|D=d,U)P(U) \quad (5)$$

We refer to this value as the exact ACE.

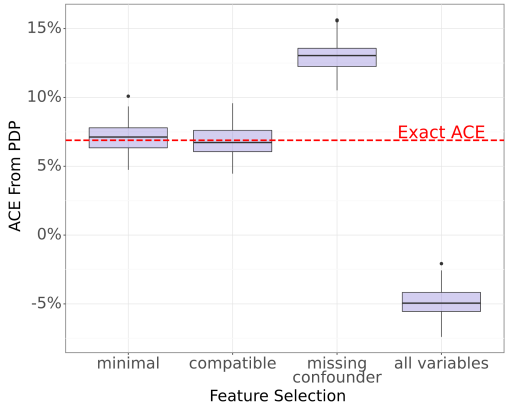

Figure 6: Average Effect of an Intervention using PDP method for different feature selections. Exact ACE is computed using do-calculus.

As previously discussed, the backdoor adjustment can be estimated from a sample population using a predictive model trained with an off-the-shelf algorithm such as XGBoost. The calculation involves a Monte-Carlo integration over a sample population of size N.

$$\begin{aligned} P(R|do(D=d)) &\simeq \frac{1}{N}\sum_{i=1}^{N} P(R|D=d,U) \\ &\simeq \frac{1}{N}\sum_{i=1}^{N} f(D=d,U) \end{aligned}$$

f is a classifier model trained to estimate the probability of *Renewal* conditional on *Discount* and *Usage*. f is applied to a sample population, taking *Usage* from the data and forcing *Discount* to the value d, as per the PDP technique.

We then compare the exact ACE with estimates from 100 sample populations of size $N = 10\,000$. For each sample population, we trained four predictive models involving different selections of features:

- *minimal*: a minimal set of features that satisfies the backdoor criterion, here $\{Discount, Usage\}$,
- *compatible*: a larger set of features compatible with the backdoor criterion, adding $\{Customer\ Profile\}$ to the minimal set,
- *missing confounder*: a set of features that does not satisfy the backdoor criterion because it excludes a variable needed to block a path between the action and the target, here excluding *Usage* from the *compatible* set,
- *all variables*: the set of all known features, incompatible with the backdoor criterion because it contains a consequence of the action, namely *Visits*.

The PDP technique is then applied to estimate the average effect on predictions of an intervention from *Discount=0* to *Discount=1*.

Figure 6 shows the experimental results. Both the *minimal* and *compatible* feature selections provide an accurate estimate of the Average Causal Effect for *Discount*. On the other hand, the two feature selections that are incompatible with the backdoor criterion lead to significantly different estimates. The calculation from the model with a missing confounder overestimates the causal effect. It is worth mentioning that here, with the classical usage of the whole set of known features, the estimate of the ACE is reversed.

6 Quantifying the intervention in a given context

Another relevant causal question is to estimate the effect of an intervention in a specific context. For an intervention on a binary variable X knowing a setting defined by the set of variables Z, the problem is to estimate an uplift from observational data (Rubin 1974; Gutierrez and Gérardy 2017):

$$uplift = P(Y|do(X=1),Z) - P(Y|do(X=0),Z) \quad (6)$$

6.1 Exact uplift using do-calculus

According to rule 2 (action/observation exchange) of the do-calculus (Pearl 2012):

$$\begin{aligned} &P(Y|do(T_1),do(T_2),K) = \\ &\quad P(Y|do(T_1),T_2,K) \text{ if } (Y \perp\!\!\!\perp T_2|T_1,K)_{G_{\overline{T_1}\underline{T_2}}} \end{aligned} \quad (7)$$

where $G_{\overline{T_1}\underline{T_2}}$ is the causal graph obtained by removing all arrows pointing to nodes in T_1 and all arrows emerging from nodes in T_2. By substituting (T_1, T_2, K) with $(\emptyset, X, Z)$:
Property *(Estimation of the effect of intervention):*

$$P(Y|do(X),Z) = P(Y|X,Z) \text{ if } (Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}}} \quad (8)$$

where $G_{\underline{X}}$ is the causal graph obtained by removing all arrows emerging from X.

In particular, if Z satisfies the backdoor criterion relative to the pair (X, Y), then the variables in Z block all paths connecting X to Y that contain an arrow into X, and further removing arrows emerging from X ensures that $(Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}}}$. Thus, if the set of variables Z satisfies the backdoor criterion relative to the pair (X, Y), then we can estimate the effect of an intervention on X by directly using conditional probabilities estimated from observational data: $P(Y|do(X = x), Z) = P(Y|X = x, Z)$.

Applied to our example (Figure 3), this property states that the uplift from *Discount* can be correctly estimated if X=*Discount*, Y=*Renewal* and Z verifies Property 8. As part of an uplift analysis, Z can be maximized, i.e. $Z = \{$*Economy, Usage, Customer profile*$\}$.

6.2 Estimate from a sample

In uplift modeling, the set of variables comprises the treatment (X) and its context (Z). Several techniques from uplift modeling can be applied to estimate an uplift from an observational dataset. In the "two models" approach, separate models are fitted on the control ($X = 0$) and treated ($X = 1$) sub-populations, as in Equation 9. In the "single model" approach, an estimator is trained on the full population, with the allocated treatment being part of the feature space as in Equation 10.

$$P(Y|do(X = x), Z) \simeq f_x(Z) \quad (9)$$

$$P(Y|do(X = x), Z) \simeq f(X = x, Z) \quad (10)$$

Subsection 6.1 demonstrated that the estimates from Equations 9 and 10 are relevant if the Property 8 holds.

6.3 Illustration: uplift from *Discount*

Since we know the causal model of our synthetic data generation process, we can compute the exact uplift from Equation 6 using pyAgrum. On the other hand, Property 8 provides an estimate from a predictive model trained on a sample population:

$$P(R|do(D), U, C, E) \simeq f(D, U, C, E) \quad (11)$$

Figure 7 compares the exact uplift calculated on the data generation causal model, with estimates from 100 classification models trained on sample populations of 50 000 observations. The left plot represents uplifts for *Corporate Customers*, and the right plot is about *Residential Customers*. The blue dots represent the exact uplift. Boxplots represent the predicted uplifts, in green for the correct selection of features (D, U, C, E) and in red for a selection comprising all known features. We observe that for *Corporate Customers*, the estimated uplifts using the correct selection of features are close to 0, regardless of the *Usage*. This is in line with the ground truth where the uplift is precisely zero. The estimated uplifts for *Private Customer* also align with the causal data generation model. However, with classical predictive approaches using the full set of known features, the uplift

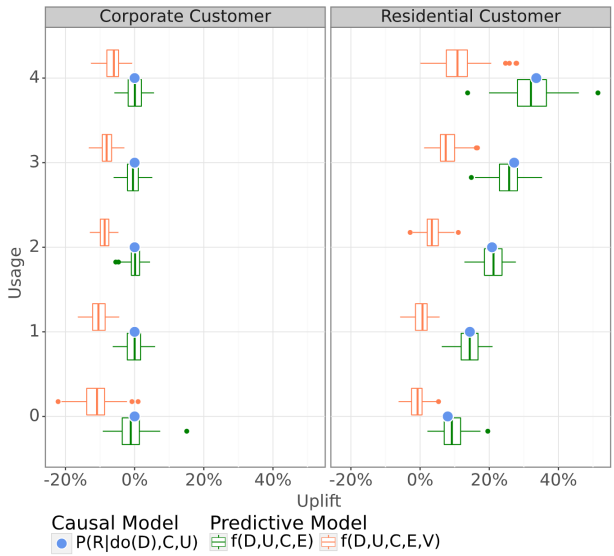

Figure 7: Predicted and theoretical uplift from an intervention on *Discount*.

estimates are far from the exact values and can even be reversed. Once again, causal effects estimated from a predictive model are quite accurate as long as the features have been carefully (and causally) selected.

7 Conclusion

The main contribution of this paper is a new XAI approach that allows better quantification of causal effects from observational data. We show that bluntly applying XAI tools to a model trained from all known features without considering causality can lead to flawed interpretations. To tackle those issues, we propose a new framework to analyze each causal query separately based on the causal structure. This leads to a tailored model and interpretability technique, providing numerical estimates of causal effects. A counterpart is that answering multiple causal queries may require training several predictive models.

In the XAI community, a debate exists around the notions *true-to-the-model* and *true-to-the-data* (Chen et al. 2020). From this perspective, it seems to us that relaxing the constraint of a pre-existing predictive model opens up the possibility for XAI to be more faithful *to-the-data*.

In our approach, causality guides both the construction and the analysis of predictive models. However, finding the complete causal structure is difficult (if possible). Different methods can be used to find a partially directed causal graph (PDAG). They can be divided into two families: methods based on conditional independence (Spirtes et al. 2002; Louis et al. 2017; Glymour, Zhang, and Spirtes 2019), and methods based on score-based methods (Chickering 2002). Hence, the next step for making our approach more operational would be to investigate how such a partial knowledge of the causal graph may be sufficient to guide predictive modeling and accurately answer causal queries.

References

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Burkart, N., and Huber, M. F. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70:245–317.

Chen, T., and Guestrin, C. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. New York, NY, USA: ACM.

Chen, H.; Janizek, J. D.; Lundberg, S.; and Lee, S.-I. 2020. True to the model or true to the data?

Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of machine learning research* 3(Nov):507–554.

Devriendt, F.; Moldovan, D.; and Verbeke, W. 2018. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. *Big Data* 6(1):13–41.

Ducamp, G.; Gonzales, C.; and Wuillemin, P.-H. 2020. aGrUM/pyAgrum : a Toolbox to Build Models and Algorithms for Probabilistic Graphical Models in Python. In *10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, 609–612.

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

Frye, C.; Rowat, C.; and Feige, I. 2020. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1229–1239. Curran Associates, Inc.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in Genetics* 10.

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

Gutierrez, P., and Gérardy, J.-Y. 2017. Causal inference and uplift modelling: A review of the literature. In Hardgrove, C.; Dorard, L.; Thompson, K.; and Douetteau, F., eds., *Proceedings of The 3rd International Conference on Predictive Applications and APIs*, volume 67 of *Proceedings of Machine Learning Research*, 1–13. PMLR.

Heskes, T.; Sijben, E.; Bucur, I. G.; and Claassen, T. 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models.

Janzing, D.; Balduzzi, D.; Grosse-Wentrup, M.; and Schölkopf, B. 2013. Quantifying causal influences. *The Annals of Statistics* 41(5):2324 – 2358.

Janzing, D.; Minorics, L.; and Bloebaum, P. 2020. Feature relevance quantification in explainable ai: A causal problem. In Chiappa, S., and Calandra, R., eds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 2907–2916. PMLR.

Kolpaczki, P.; Bengs, V.; and Hüllermeier, E. 2023. Approximating the shapley value without marginal contributions.

Louis, V.; Sella, N.; Affeldt, S.; Singh, P. P.; and Isambert, H. 2017. Learning causal networks with latent variables from multivariate information in genomic data. *Public Library of Science Computational Biology* 13.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lundberg, S. M., and Lee, S.-I. 2018. Shap. https://github.com/slundberg/shap.

Pearl, J. 2000. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pearl, J. 2012. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, 3–11. Virginia, USA: AUAI Press.

Rieg, T.; Frick, J.; Baumgartl, H.; and Buettner, R. 2020. Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms. *PLOS ONE* 15(12):1–20.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.

Shapley, L. S. 1953. A value for n-person games. In Kuhn, H. W., and Tucker, A. W., eds., *Contributions to the Theory of Games II*. Princeton University Press. 307–317.

Spirtes, P.; Glymour, C.; Scheines, R.; Kauffman, S.; Aimale, V.; and Wimberly, F. 2002. Constructing bayesian network models of gene expression networks from microarray data. *Proc. of the Atlantic Symposium on Computational Biology, Genome Information Systems & Technology*.

Strumbelj, E., and Kononenko, I. 2010. An efficient explanation of individual classifications using game theory. *Journal Of Machine Learning Research* 11:1–18.

Sundararajan, M., and Najmi, A. 2020. The many shapley values for model explanation. In III, H. D., and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 9269–9278.

von Neumann, J., and Morgenstern, O. 1947. *Theory of games and economic behavior*. Princeton University Press.

Wang, J.; Wiens, J.; and Lundberg, S. 2021. Shapley flow: A graph-based approach to interpreting model predictions. In Banerjee, A., and Fukumizu, K., eds., *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 721–729. PMLR.

Zhao, Q., and Hastie, T. 2019. Causal interpretations of black-box models. *Journal of business and economic statistics : a publication of the American Statistical Association* 2019.