Reliable Estimation of Causal Effects Using Predictive Models

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

Reliable Estimation of Causal Effects using Predictive Models

Mahdi HADJ ALI Yann LE BIANNIC
Pierre-Henri WUILLEMIN
LIP6 UMR 7606 Sorbonne Université - CNRS

June 14, 2024

Abstract

In recent years, machine learning algorithms have been widely adopted across many fields due to their efficiency and versatility. 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 predictions of a pre-trained model. Drawing on these improvements, practitioners seek to gain causal insights into the underlying data-generating mechanisms. To this end, works have attempted to integrate causal knowledge into interpretability, as non-causal techniques can lead to paradoxical explanations. In this paper, we argue that each question about a causal effect requires its own reasoning and that relying on an initial predictive model trained on an arbitrary set of variables may result in quantification problems when estimating all possible effects. As an alternative, we advocate for a query-driven methodology that addresses each causal question separately. Assuming that the causal structure relating the variables is known, we propose to employ the tools of causal inference to quantify a particular effect as a formula involving observable probabilities. We then derive conditions on the selection of variables to train a predictive model that is tailored for the causal question of interest. Finally, we identify suitable eXplainable AI (XAI) techniques to estimate causal effects from the model predictions. Furthermore, we introduce a novel method for estimating direct effects through intervention on causal mechanisms.

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. Interpretability is crucial in various fields, such as medical prescription or legal domain [34, 9]. Indeed, using models for automated decision-making entails understanding their behavior to justify choices.

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[28] 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 solely on assumptions about the structure of the causal model for the identification of causal effects and to involve additional predictive models learned from data for numerical estimates.

Various works deal with quantifying causal effects (direct and/or indirect) from a predictive model, presuming knowledge of the causal structure that depicts the relations among features [19, 43]. These studies fall within the current focus of the eXplainable AI (XAI) field [2]: their starting point is a predictive model, typically trained from all known variables for a classification or regression task, and their purpose is to explain how each input feature contributes to the model predictions. The objective of this paper is to show the benefits of an alternative approach, wherein a predictive model is no longer pre-existing but is rather tailored to address a specific causal query.

As in previous works [13, 19, 43, 45], the paper assumes prior knowledge about the causal structure, but we propose using it before building, training, and analyzing query-driven predictive models from observational data.

This paper introduces an innovative methodology for quantifying a total causal effect using a predictive model and estimating a direct causal impact using a novel extension of the causal structure.

The first section explores explainability paradigms and tools for predictive models. The second discusses the limitations of predefined models. The third introduces a new methodology for accurate causal effect estimation. Finally, the last section presents our new framework for interventions on causal mechanisms.

2 Different explainability paradigms

Apart from predictive ability, there is a growing interest in the capacity to explain predictions. Transparent white-box models reveal their inner mechanisms in human-readable forms. So graphical and causal models, especially in classification, are intuitive white boxes that facilitate understanding

through visuals. However, the accuracy-explainability trade-off may favor complex black-box models, effective but less transparent to humans [17], thus motivating dedicated explanation tools [3].

In this section, we will briefly introduce the white-box and black-box approaches, from the explanation perspective.

2.1 Graphical models, causal models

Graphical and causal models visually show relationships between variables, aiding human understanding of influences and, thus, explaining the decision-making process. Explaining a prediction for humans means analyzing its causes, a straightforward task for a graphical model with causal semantics (causal model).

2.1.1 Graphical Models

A Bayesian network, also known as a probabilistic graphical model, represents probabilistic relationships among a set of variables. The variables are represented as nodes, and the relationships between variables are represented as directed edges between the nodes. Each node in the network corresponds to a random variable, and the edges indicate the (direct) probabilistic dependencies between variables. Its structure is based on a directed acyclic graph (DAG), meaning the edges form a directed flow without any cycles. This structure encodes that each variable is conditionally independent of its non-descendants given its parents (Local Markov Property), leading to a factorization of the joint distribution $\mathbb{P}$ of the N variables in the model :

$$\mathbb{P}(X_1, \cdots, X_N) = \prod_{i=1}^{N} \mathbb{P}(X_i | Pa(X_i)) \tag{1}$$

where $Pa(X_i)$ is the set of parents of X_i in the DAG.

Definition 2.1 *(Markov Compatibility) [28]*
If a joint distribution $\mathbb{P}$ admits the factorization of Eq.1 compatible to a DAG $\mathcal{G}$, we say that $\mathcal{G}$ and $\mathbb{P}$ are compatible and that $\mathbb{P}$ is Markov relative to $\mathcal{G}$.

A practical way to describe the set of distributions compatible with a DAG $\mathcal{G}$ is to list the set of conditional independencies each distribution must satisfy. These independencies can be read from $\mathcal{G}$ via a graphical criterion called d-Separation [28]. The concept of d-separation allows us to identify conditional independence relationships in Bayesian networks. Thus connecting probability and graphical considerations. If X is d-separated from Y given Z, then X and Y are conditionally independent given Z. In other words, knowing the values of the nodes in set Z "blocks" the information flow between X and Y, making them independent of each other. To determine whether two sets of nodes X and Y are d-separated given a set of nodes Z, Pearl in [28] established two criterions:

Definition 2.2 *(Blocked path)*
A path p is said to be blocked by a set of nodes Z if and only if :

(i) either p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z,

(ii) or p contains a collider $i \to m \leftarrow j$ such that Z does not contain m or one of its descendants.

Definition 2.3 *(d-Separation)*
A set Z is said to d-separate the sets X and Y if and only if Z blocks every path from a node in X to a node in Y.

Building on the insights from d-separation, we introduce the Markov boundary of a target variable [27]. It represents the minimal set of variables that renders the target conditionally independent of all other variables. In other words, the Markov boundary contains the minimal set of variables that are necessary for predicting the target. Grasping the Markov boundary of a target variable helps specify key influential features for predictions. This insight into the directly affecting variables enhances explanation clarity and asserts Bayesian networks as white-box models.

2.1.2 Causal Models

In [28, 31], Pearl proposes a complete causal framework based on a probabilistic graphical model or structural equation model [30]. Unlike correlation, which only measures the strength of statistical dependencies, Pearl's causality aims to uncover and quantify the underlying mechanisms that generate these relationships. In the subsequent discussion, we will distinguish between the "causal structure" (i.e. the graphical representation of causal relations) and the "causal model" (i.e. the causal structure along with its learnable parameters derived from observations).

The causal graphical model is often represented using DAGs as Bayesian Networks. However, Bayesian networks represent probabilistic relationships between variables, whereas causal graphs explicitly represent causal relationships, allowing for an intuitive understanding of the cause-effect structure among variables. The directed edges in the graph indicate causal relationships, and the absence of an edge implies the absence of a direct causal link. Unlike Bayesian networks, a variable can be unobserved but still be in the causal model (latent variables).

Another method for illustrating causal relationships is found in the framework of Functional Causal Models (*FCM*). These models express causal links using deterministic functional equations, and the incorporation of probabilities arises from the assumption that certain variables within these equations remain unobserved. A Functional Causal Model comprises a series of equations formulated as follows.

Definition 2.4 *(Pearl's Functional Causal Model)*
In [30], Pearl defines a functional causal model as:

(1) A set $\mathfrak{U} = \{U_1, \cdots, U_N\}$ *of background or exogenous variables, representing factors outside the model, which still affect relations within the model.*

(2) A set $\mathcal{V} = \{V_1, \cdots, V_N\}$ *of observed endogenous variables, where each* V_i *is functionally dependent on a subset* $Pa(v_i)$ *of* $\mathfrak{U} \cup \mathcal{V} \setminus \{V_i\}$.

(3) A set $\mathcal{F}$ *of functions* $\{F_{V_1}, \cdots, F_{V_N}\}$ *such that each* F_{V_i} *determines the value* v_i *of* $V_i \in \mathcal{V}$, $v_i = F_{V_i}(Pa(v_i), u_i)$

(4) A joint probability distribution $\mathbb{P}(u)$ *over* $\mathfrak{U}$.

A FCM Ψ is then written as a set of functional equations :

$$\Psi : \begin{cases} V_1 & = F_{V_1}(Pa(V_1), u_1) \\ & \cdots \\ V_N & = F_{V_N}(Pa(V_N), u_N) \end{cases} \tag{2}$$

It is important to note that no assumptions are made regarding the nature of the distribution for $\mathfrak{U}$. However, under assumptions of acyclicity and joint independencies of $\mathfrak{U}$, Pearl and Verma in [32] and then Druzdzel and Simon [10] established that Bayesian networks and functional models share an essential equivalence in probabilistic modeling. They showed that any Bayesian network $\mathcal{G}$ on $\mathcal{V}$, characterized by a distribution $\mathbb{P}(\mathcal{V})$ (as in Eq.1), can be associated with an equivalent *FCM* (as in Def 2.4). This model, in turn, can generate a distribution identical to $\mathbb{P}$. This equivalence illustrates the smooth interchangeability of probabilistic applications with Bayesian networks, including statistical estimation, prediction, and diagnosis, to functional models, and vice versa, and highlights the value of functional models as an alternative representation of joint distribution functions, providing a dual perspective that enhances reasoning and analysis across both modelings.

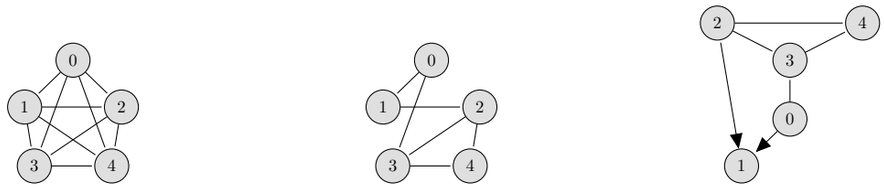
2.1.3 Causal Discovery

Despite employing careful statistical and computational techniques, certain aspects of a causal model remain unattainable to automatic learning from observations. For instance, latent variables, confounders, symmetrical relations between pairs of variables, and other factors can complicate the learning process.

For Bayesian networks, there exist several algorithms for structural learning [15] which fall into two categories: constraint-based methods and score-based methods. The latter scores different potential structures and chooses the one with the highest score w.r.t the data [7]. Constraint-based methods rely on the dependencies that can be tested in the data; see Fig.1. They start by locating independencies and then remove edges when independence is observed. The next step is to determine edge direction thanks to tested conditional independence or constraint propagation. Notable algorithms within this family include PC[37], FCI[36], RFCI [8], and MIIC [25]. As a distinguishing feature, MIIC sets itself apart by basing its independence tests on the principle of mutual information.

As depicted in Fig.1 Step 3, these algorithms may end in an incomplete DAG known as a Partial Directed Acyclic Graph (PDAG). This mixed structure represents the Markov-equivalence class, containing all Bayesian networks achieved by assigning directions to non-oriented edges.

Assuming that within this class, only one structure represents the causal model, the PDAG also acts as the partial causal graph inferred from available data. However, it's essential to acknowledge that this assumption overlooks factors like latent variables that influence causal relationships.



(Step1) Fully connected graph **(Step2)** Independencies removal **(Step3)** Orientations

Figure 1: Steps for constraint-based algorithm

2.1.4 A common causal query: the Average Causal Effect

Learning a causal structure supports qualitative causal insights about a target variable Y, such as identifying variables with a direct or indirect causal effect on Y. However, quantitative insights about

the effect of a specific variable on the target Y involve numerical estimates, such as the Average Causal Effect (*ACE*) [20, 28, 21]. For a binary variable X, the *ACE* is defined as the difference of the predictions for the target Y when one forces the value of X to 1 or 0.

Definition 2.5 *(ACE)*

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

When a causal graphical model is applied to a classification task, computing the *ACE* for each known variable would be a step towards explaining the model. However, there are several practical limitations to this approach. Discovering a complete causal graph is a hard problem, as discussed in Subsection 2.1.3. The existence of latent variables, or numerous parents for some nodes, may further compound the estimation of conditional probabilities involved in do-calculus formulas.

On the other hand, the statistical learning theory [41] has led to effective algorithms and models to accurately estimate a dependent variable from data without any prior knowledge.

2.2 Predictive models and Explainability

Classification is a common task in supervised learning, i.e. to predict a (binary) class Y of an object from a vector of features $\mathbf{X} = \{X_1, \cdots, X_j, \cdots, X_N\}$. An ML model is trained from a database of observations $\mathcal{D}$ 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 the probability of the target class:

$$f(\mathbf{X}) \simeq \mathbb{P}(Y=1|\mathbf{X})$$

While there exist numerous other explainability methods [2, 40], we will focus on two techniques: Partial Dependence Plot and Shapley values, because of their extensive usage and the existing research on linking them to causal knowledge.

2.2.1 Partial Dependence Plots

A *Partial Dependence Plot*(*PDP*) displays the marginal effect that a subset S (of input features) has on the output of a predictive model [12]. The partial dependency between the model output and the vector of features $\mathbf{X_S}$ is estimated by marginalizing the predictions over the remaining input features:

$$\forall S \subset [\![1, \cdots, N]\!], PDP(\mathbf{X_S}) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} f(\mathbf{X_S}, \mathbf{x_{\bar{S}}^{(i)}}) \tag{3}$$

where $\bar{S} = [\![1, \cdots, N]\!] \setminus S$, $\mathbf{X_S} = \{X_j, j \in S\}$ and $\mathbf{x_{\bar{S}}^{(i)}} = \{x_j^{(i)}, j \in \bar{S}\}$ with $x_j^{(i)}$ being the value of the j-th variable in the i-th sample in the database $\mathcal{D}$.

A *PDP* offers valuable insights into dependencies between model predictions and a subset of input features. Each *PDP* is typically computed per feature to gauge strength, linearity, nonlinearity, monotonicity, or threshold influence, etc.

The *Individual Conditional Expectation Plot* (*ICE*) [16] complements *PDP* by assessing whether a feature's contribution to predictions hinges on interactions with other features. It also visually verifies if the function implemented by the model is additive. In contrast to a *PDP* which illustrates the average marginal effect of an input feature on predictions, each line in an *ICE* plot represents the marginal effect for a sample observation when varying the feature of interest and fixing the other features to their observed value.

$$\forall S \subset [\![1, \cdots, N]\!], \begin{cases} \forall i \in [\![1, \cdots, |\mathcal{D}|]\!], & ICE^{(i)}(\mathbf{X_S}) = f\left(\mathbf{X_S}, \mathbf{x_{\bar{S}}^{(i)}}\right) \\ & PDP(\mathbf{X_S}) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} ICE^{(i)}(\mathbf{X_S}) \end{cases} \tag{4}$$

Parallel *ICE* lines suggest an additive feature contribution without evident feature interactions. Alternatively, interaction analysis can be eased by centering *ICE*, aligning all lines through a given point in the plot.

PDP and *ICE* plots depict a model's reliance on an input feature X. Dependency indicates that the model uses X to predict the target Y, yet doesn't necessarily imply that X causally affects Y in data generation. For instance, information could come from any Y-correlated variable, even a causal consequence of Y.

Zhao and Hastie [45] observed that Equation 11 is a Monte-Carlo approximation of Pearl's backdoor adjustment (see below section 4.2.2). They identified three requirements for the *PDP* of a feature to support a successful causal interpretation: (i) an *accurate predictive model* closely approximating conditional probabilities from the data generation process, (ii) *domain knowledge* about the causal structure to ensure that Pearl's back-door condition is satisfied by the remaining model inputs, and (iii) *visualization tools* like *PDP* and *ICE* to spot unexpected behaviors from unmeasured confounding or flawed causal structure assumptions.

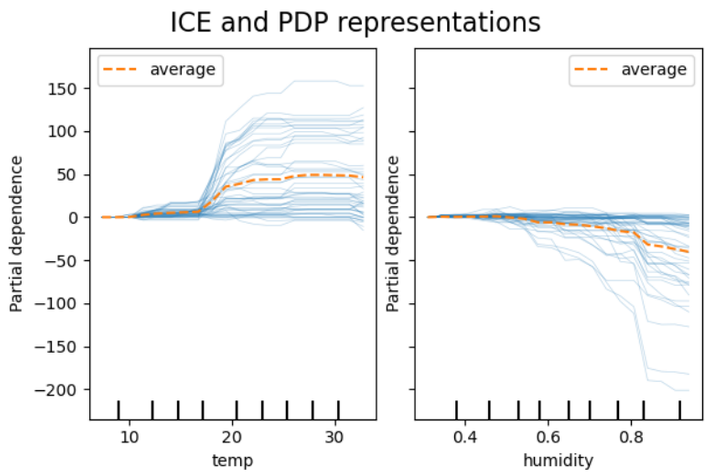


Figure 2: Partial Dependence Plot (named average) represented as a dashed orange line and Individual Conditional Expectation Plots drawn as solid blue lines, an example from scikit-learn[33].

2.2.2 Shapley values

Another extensively employed XAI technique is derived from game theory: Shapley values [35] are a method to spread credit among a set of players $\mathbf{X}$ in a coalition game [42]. It is a "fair" attribution in the sense that it rewards each player according to his contribution, and it is the unique solution that satisfies four desirable properties: efficiency, linearity, symmetry, and nullity, as well as other properties such as monotonicity[44].

In this framework, a value function v associates a real number $v(S)$ to any coalition $S \subseteq \mathbf{X}$ of players. $v(S)$ represents the total expected payoff that the members of S can obtain by cooperating. Shapley values are estimated by imagining that coalitions grow incrementally, one player at a time. Each player entering a previously formed coalition S may then demand a fair compensation for its expected marginal contribution $v(S \cup \{i\}) - v(S)$. A player's contribution is estimated by averaging their marginal contribution over all possible coalition permutations.

The Shapley value for the variable X_i is then :

$$\phi_i = \sum_{S \subseteq \mathbf{X} \setminus \{X_i\}} \frac{|S|!(N - |S| - 1)!}{N!} \Big(v(S \cup \{X_i\}) - v(S) \Big) \tag{5}$$

To transpose this framework to XAI, a parallel is drawn between a model's input features $\mathbf{X}$ and the players who cooperate in a game [38]. Shapley values can then provide a solution, backed by a solid theory, to the feature attributions problem. A challenge for applying Shapley values to ML is the definition, from the model, of a coalition game and its value function v. When the coalition is the full set of features, a natural value function is the prediction made by the model. For a subset of features, the influence of missing features on model predictions can be assessed through several variants, such as fixing the values in S and estimating the expected model prediction over some distribution of the remaining features. Since each variant defines a specific coalition game, the corresponding Shapley values are named after the variant. Two popular variants in the literature are conditional Shapley values [1] and marginal Shapley values [39]:

$$v^{cond}(S) = \mathbb{E}\big[f(X_{\overline{S}}, x_S) | X_S = x_S\big] \tag{6}$$

$$v^{marg}(S) = \mathbb{E}\big[f(X_{\overline{S}}, x_S)\big] \tag{7}$$

Both variants present mathematical issues, as discussed in [24] and [4]. Conditional Shapley values tend to spread credit between correlated features, even redundant features discarded by the learning algorithm. Moreover, their exact calculation involves modeling an exponential number of multivariate distributions. On the other hand, marginal Shapley values may require evaluating the model on "out-of-distribution" samples, risking extrapolation into unobserved or implausible parts of the feature space.

Causality and Shapley values

Janzing et al. in [22] present a causal perspective on Shapley values by substituting observational conditioning with intervention-based conditioning as in Pearl's do-calculus [28]. They argue that a formal distinction can be made between the algorithm's inputs and the real-world context features when seeking causal insights. Consequently, the authors propose that *Interventional Shapley values* reduce to *Marginal Shapley values*, thereby justifying the use of the latest for causal inquiry. This proposal explains the causal mechanism linking model inputs to outputs and is agnostic about the causal relations between real-world features.

Other authors aimed at a different goal: providing an explanation that considers real-world causal relations so that credit may be attributed appropriately to root causes. In line with this goal, Frye et al. introduced a method to incorporate causality into Shapley values, called *Asymmetric Shapley values* [13]. The reasoning is as follows: if X_i is known to be the deterministic causal ancestor of

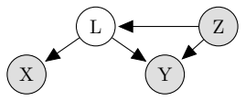


Figure 3: Variable X has predictive power about Y but no causal effect

X_j, then we might want to attribute all the predictive contributions to X_i and none to X_j. To achieve this, the authors only consider permutations that are consistent with a partial causal order. Moreover, they consider that unconditional marginalization may extrapolate outside the data domain. In agreement with [1], they propose to use observational conditioning while introducing techniques to handle high-dimensional data.

Heskes et al. propose to consider a more detailed causal structure in the form of a causal chain graph, and to use Pearl's do-calculus to estimate the real-world effect of interventions on in-coalition features. The characteristic function for these *Causal Shapley values* [19] is:

$$v^{caus}(S) = \mathbb{E}\big[f(X_{\overline{S}}, x_S)|do(X_S = x_S)\big]$$

For each permutation, the authors identify the contribution of a feature as a total causal effect and show that it can then be decomposed into a direct and indirect effect.

We remark that an arbitrary predictive model may have been trained on features that are neither direct nor indirect causes of the target; such features can still offer the model non-redundant information about the target. For instance, when a variable X has no causal descendant and is a direct consequence of a latent cause L of the target Y (see fig. 3), X has unique predictive power about Y. Since X has no outgoing arc, intervening on X will not change the distribution of other variables. However, the intervention will influence the model predictions in v^{caus}, and may thus obfuscate the causal interpretation of other variables (e.g. Z).

In [43], Wang et al. observe that a Shapley value depends on the other variables included in the model and that credit for causal effects is divided among upstream and downstream variables in a causal path. They develop Shapley Flows by reformulating the problem to assign credit to edges rather than nodes in a causal graph. Shapley Flows highlight an issue that is common to all causal interpretations of predictive models: given a model and the causal structure of its variables, a root cause input variable that is outside the Markov boundary of the target may have been ignored by the model learning algorithm, either implicitly or during an explicit feature selection step [14]. Estimating the total causal effect of such a variable must then involve additional statistical models (in the case of Shapley Flows, extra linear regression or gradient boosting models).

The following section introduces an experimental protocol for formally calculating ground truth causal effects. Then, it illustrates how standard XAI techniques are sensitive to feature selection prior to training a predictive model and may lead to paradoxical explanations.

3 Paradoxical insights from XAI

In practice, models learned from datasets may be biased. To overcome these problems, we propose a study based on an experimental setting.

3.1 Experimental Protocol

We propose a causal Bayesian network as the ground truth reference. We designed a synthetic example with pyAgrum, a library for probabilistic graphical models [11]. 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:

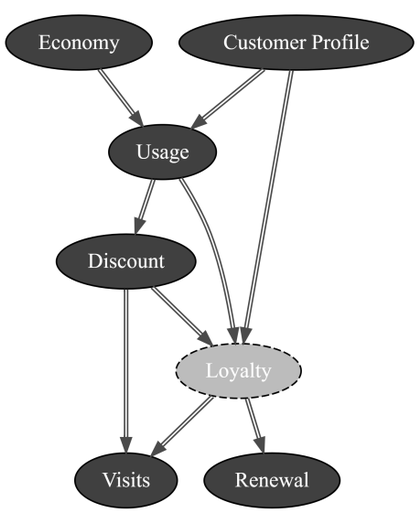


Figure 4: The causal Bayesian that generates the dataset.

- *Economy* (noted as E) represents economics conditions
- the client profile (e.g. residential vs commercial) is represented by the variable *Customer Profile* (noted as C),
- the yearly consumption of the customer is tracked by *Usage* (noted U)
- an offer granted to the client 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 4 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*):

$$\begin{cases} \mathbb{P}(R|do(D=1), C=0) > \mathbb{P}(R|do(D=0), C=0) \\ \mathbb{P}(R|do(D=1), C=1) = \mathbb{P}(R|do(D=0), C=1) \\ \mathbb{P}(R|do(D=1)) > \mathbb{P}(R|do(D=0)) \end{cases}$$

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

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 [28]. Thus, we can examine a classification model's interpretations and assess their consistency with the underlying causal model.

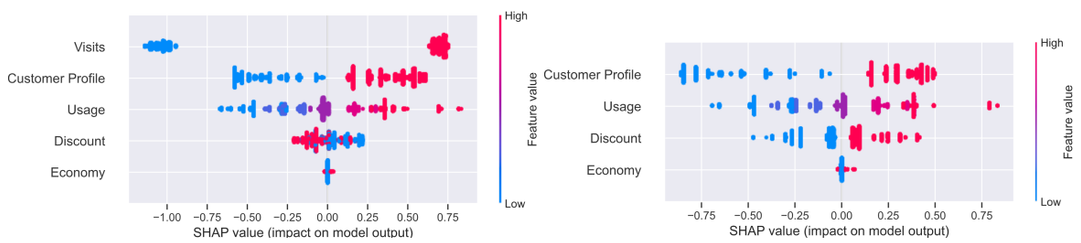
The purpose of the next sections is to show in different contexts how the causal interpretation of classical XAI results can be ambiguous (section 3.2) and how our proposition can lead to more consistent estimations of causal effects from specific prediction tasks.

3.2 Sensitivity to feature selection

It is known that feature selection has a significant impact in predictive modeling [23]. This section illustrates how selecting features without causal analysis can lead to paradoxes when applying common XAI techniques to the learned model.

3.2.1 Paradoxical insights from SHAP

We used the well-known XGBoost algorithm [6] to train two models on the same dataset but using different sets of features. A first model was trained on all known features (i.e., all variables except the target and the unobserved *Loyalty*), and a second model was trained after dropping *Visits*. For these two models, we used the SHAP library [26] to compute the marginal Shapley values of the features. The results are given in Figure 5a and Figure 5b.



(a) Explanation of a model on all variables. **(b)** Explanation of a model excluding *Visits*.

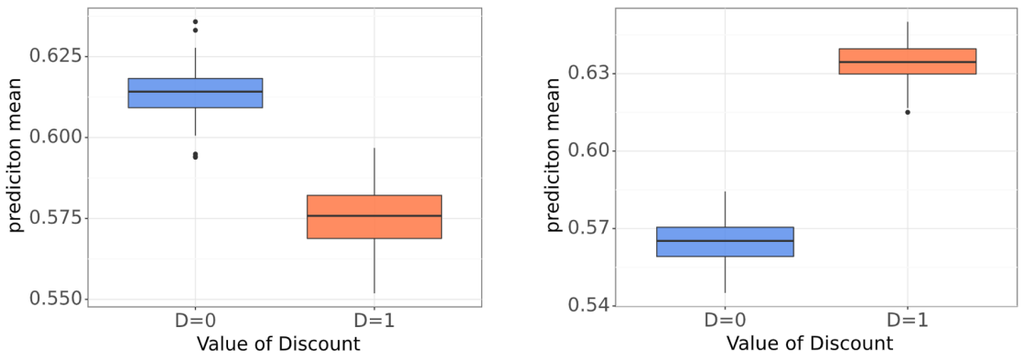
Figure 5: Summary Plot from SHAP

The two plots display the marginal Shapley values of every feature for a sample population, as documented in the SHAP library. Each dot represents the marginal Shapley value of a feature for a sample observation, in the log-odds space. The dot 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 predictions in the first model Figure 5a, while it has a positive contribution in the second model Figure 5b. If marginal Shapley values were naively interpreted as direct causal effects happening in the data generation process, an analyst might draw opposite conclusions from the two models.

3.2.2 PDP sensitivity to feature selection

In the previous section, we used marginal Shapley values that are considered "true to the model" in literature [22], [4]. To address the risk that the paradox could be an artifact of the learning and interpreting pipeline, we replicated the experiment on 100 different populations of 10k observations generated from the same data generation process, and used the Partial Dependence Plot technique that relies on simpler marginal expectations (the *PDP* formula for X is indeed equivalent to $v(X)$ in marginal Shapley calculations). In Figure 6, we present the *PDP* results for the variable *Discount*, demonstrating its influence on predictions for the two variable selections. Each boxplot represents 100 results obtained through the *PDP* analysis. Upon initial observation, it becomes evident that the results do not exhibit qualitative consistency. The right plot demonstrates a positive impact, while the other indicates a negative impact on the model's predictions.



(a) Explanation of a model on all variables.

(b) Explanation of a model excluding *Visits*.

Figure 6: PDP of Discount on Renewal

Transitioning from the sensitivity of SHAP to feature selection to that of *PDP* it is crucial to acknowledge the significant impact variable selection can have on the interpretability of partial dependence plots. When including irrelevant variables or omitting essential features, the plot's reliability, and accuracy may diminish, leading to misinterpretations.

3.2.3 Discussion

We observed that the widespread SHAP and *PDP* interpretation methods are sensitive to feature selection: they provide conflicting insights when applied to different models trained using the same ML algorithm, on the same population, but with different selections of features. An analyst may thus wonder whether the model explanations are consistent with real-world causal effects.

Several authors have proposed incorporating knowledge about the causal structure when interpreting a given predictive model. However, the quantification of causal effects may require information that cannot be extracted from the model. For instance, a root cause (e.g. *Economy*) may be outside the Markov boundary of the target in the causal graph of the model features, and may thus be ignored during the learning process. On the other hand, a variable that is not a causal ancestor of the target (e.g. *Visits*) may have a strong predictive power that obfuscates the interpretation of other variables. Therefore, in order to assess the real-world causal effect of a specific input variable from an arbitrary predictive model, an XAI technique may not get all relevant information from the model itself and may depend on additional probabilistic models.

Thus, we argue that a predictive model that is typically optimized for accurate prediction of a target from all available information is not the ideal starting point to analyze a causal effect involving a specific variable. This motivates the methodology that we propose in the next section to address precise causal queries.

4 Towards a hybrid methodology

4.1 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 causal relationships of variables [13, 19]. 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 7.

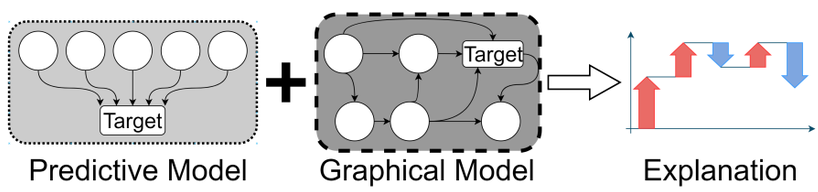


Figure 7: Classic XAI Approach.

In this paper, we propose a new methodology, illustrated in Figure 8, which extends the common framework previously described. Our approach consists of several phases. First, we start with a training population, a causal graph, and a specific query about a causal effect. We then use the tools of causal inference to quantify the causal effect in terms of observable probabilities. Next, we train an ML model to estimate these probabilities. Finally, we use an interpretability method to compute the requested causal effect from the model predictions.

A key difference between our proposal and previous methods is that we do not assume a pre-trained model. The main argument is that a single general predictive model cannot systematically answer different causal questions.

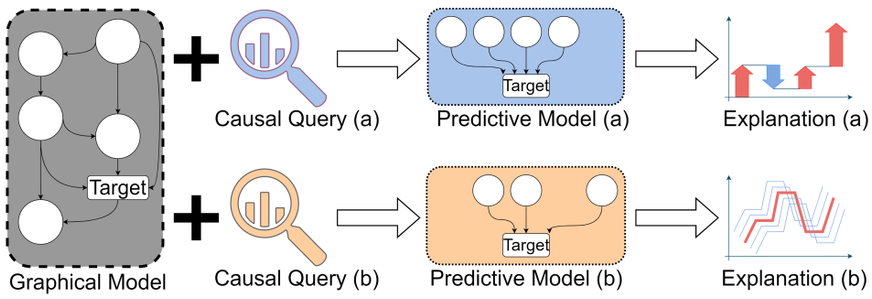


Figure 8: Proposed approach for two distinct queries.

4.2 Methodology for estimating a total causal effect

With the methodology now introduced, the approach can be utilized to answer various causal queries. Let us assume that the objective is to analyze the effect of a discount on subscriber renewal.

4.2.1 Exact solution using do-calculus

Within a probabilistic causal framework, the query for the total causal effect is the quantification of the probability $\mathbb{P}(Y|do(X))$. In such a framework, do-calculus provides multiple techniques, such as frontdoor or backdoor adjustments, to compute such causal effects [28]. In particular, the backdoor adjustment defines a set of variables that should be considered.

Definition 4.1 *(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 4.2 *(Backdoor Adjustment)*
If Z satisfies the backdoor criterion relative to (X, Y):

$$\mathbb{P}(Y|do(X = x)) = \sum_z \mathbb{P}(Y|X = x, Z = z)\mathbb{P}(Z = z) \tag{8}$$

Applied to our example (Figure 4), the backdoor adjustment is suitable for quantifying the causal effect of *Discount* on *Renewal* with $\{Usage\}$ as a set satisfying the backdoor criterion.

4.2.2 Estimates from a sample data

Estimating the causal effect through the backdoor adjustment in Equation 8 only involves the variables Y, X, and Z. Equation 8 can be generalized and reformulated using $X_S = \{X\}$, $X_{\bar{S}} = Z$:

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

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 $\mathbb{P}(Y|\mathbf{X})$ is estimated by $f(\mathbf{X})$:

$$\mathbb{P}(Y|do(x_S)) \simeq \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}(Y|X_S = x_S, X^i_{\bar{S}}) \tag{9}$$

$$\simeq \frac{1}{N}\sum_{i=1}^{N} f(x_S, X^i_{\bar{S}}). \tag{10}$$

Zhao and Hastie (2019) already demonstrated the analogy between the backdoor adjustment and the *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 11.

$$f_S(x_S) = E_{X_{\bar{S}}}[f(x_S, X_{\bar{S}})] = \int f(x_S, x_{\bar{S}})d\mathbb{P}(x_{\bar{S}}) \tag{11}$$

Indeed, the Monte-Carlo integration of Equation 11 over the training data has precisely the same equation as Equation 10. 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.

4.2.3 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 12). Indeed two sets satisfy the criterion 4.1: $\{Usage\}$ and $\{Usage, Customer\ profile\}$. For the first set Equation 8 becomes:

$$\mathbb{P}(R|do(D=d)) = \sum_{U} \mathbb{P}(R|D=d,U)\mathbb{P}(U) \tag{12}$$

We refer to this value as the exact ACE.

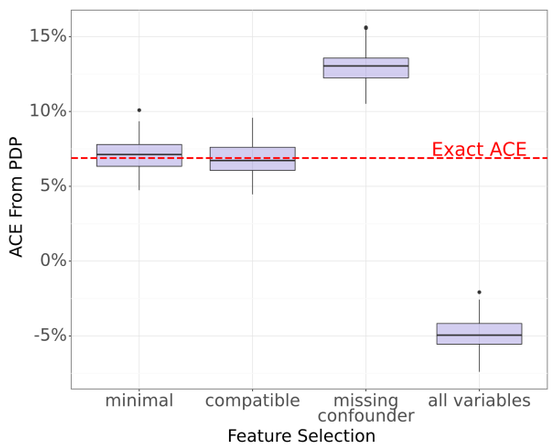


Figure 9: Average Effect of an Intervention using *PDP* method for different feature selections. Exact ACE is computed using do-calculus.

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 $|\mathcal{D}|$.

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

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*: 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 intervention from *Discount=0* to *Discount=1*.

Figure 9 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 when employing the traditional approach that utilizes the complete set of known features, the estimate *ACE* becomes inverted.

In short, starting with a causal model and a query about a variable's total causal effect, we employed do-calculus to quantify this effect using observable data. The probabilities can be estimated via a predictive model, trained on a well-chosen feature selection. In our case, the table below presents compatible feature selections for estimating the *ACE* of known variables (denoted by their first letter). It illustrates the importance of training separate models to estimate the causal effect of different variables.

ACE on R	Feature Selection for Predictive Model
E	{E} or {E, C}
C	{C} or {E, C}
U	{U,C} or {U,C,E}
D	{D,U}, {D,U,C}, {D,U,E} or {D,U,C,E}
V	impossible (confounding path)

Table 1: Feature selection that satisfies a criterion provided by do-calculus.

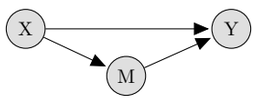


Figure 10: An unconfounded mediation model with treatment (X), mediator (M) and target (Y)

Having presented the methodology for determining the total causal effect of a variable, in the following sections, we will proceed with the same approach to explore other causal questions. Particularly, we will delve into discussions concerning direct effects. Note that additional exploration of its application for uplift analysis was detailed in [18].

5 Direct Effect

Up to this point, the study has primarily centered on total causal effects, which are the most straightforward causal relationships to comprehend, identify, and estimate. Yet, the main interest might be understanding the direct influence of a variable X on the outcome Y. In other words, the "direct causal effect" of an exposure X on an outcome Y measures how responsive Y is to changes in X while blocking for mediator effects.

5.1 Controlled Direct Effect

The Controlled Direct Effect (*CDE*) is the direct effect observed when fixing mediators to a fixed value, effectively blocking the influence of the exposure on them.

5.1.1 Exact computation of the *CDE*

Consider the mediation model in Figure 10. In this setup, the treatment X is assumed to affect the outcome Y both directly and indirectly through the mediator M. The *CDE* measures the part of the effect of X on Y that is not due to changes in M. Formally, the *CDE* of X on an outcome Y, controlling for a mediator M at level m, is defined by Pearl [29] as:

$$CDE(x, x', m) = \mathbb{P}(Y|do(X = x, M = m)) - \mathbb{P}(Y|do(X = x', M = m)) \tag{13}$$

Considering that *CDE* is a do-expression, its identification can be fully achieved utilizing the principles of do-calculus [30].

5.1.2 Illustration: Controlled Direct Effect of *Usage* on *Renewal*

In our example, we can compute the controlled direct effect of *Usage* on *Renewal* with *Discount* as a mediator: $\mathbb{P}(R|do(U = u, D = d))$ The *CDE* formula is identified accordingly to do-calculus:

$$\mathbb{P}(R|do(U = u, D = d)) = \sum_C \mathbb{P}(R|C, D = d, U = u)\mathbb{P}(C)$$

As in section 4.2.2, to compute this quantity from observational data, the proper procedure is to build a probabilistic model f of R, knowing only C, D, U. Then use a Monte Carlo integration and a population of size $|\mathcal{D}|$ and apply the model $f(.)$ to quantify the probability $\mathbb{P}(R|C, D, U)$:

$$\begin{aligned}\mathbb{P}(R|do(U = u, D = d)) &\simeq \tfrac{1}{|\mathcal{D}|}\textstyle\sum_{i=1}^{|\mathcal{D}|}\mathbb{P}(R|U = u, D = d, C = c^i)\\ &\simeq \tfrac{1}{|\mathcal{D}|}\textstyle\sum_{i=1}^{|\mathcal{D}|} f(U = u, D = d, C = c^i)\end{aligned}$$

5.2 Intervention on a causal mechanism

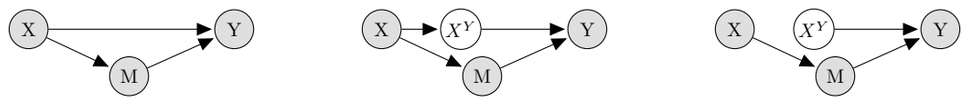
By analogy with the analysis of a total causal effect as the effect of an intervention on a variable, we propose in this section to analyze a direct causal effect by studying the effect of an intervention on a direct causal mechanism in a graphical model.

Definition 5.1 *(Intervention on a causal mechanism)*

In a FCMΨ (see Def.2.4), we define an intervention on the causal mechanism $X \to Y$ through which X influences Y as a modification of the structural equation of Y, where X becomes fixed to a value x:

$$Y = F_Y(Pa(Y) \setminus \{X\}, X, u_Y) \xrightarrow{becomes} Y = F_Y(Pa(Y) \setminus \{X\}, X = x, u_Y)$$

where F_Y is an arbitrary function, $Pa(Y)$ is the set of variables that directly determines the value of Y, and u_Y represents disturbances due to omitted factors. The remaining equations involving X are unaffected by the intervention.



(a) Original Graph $\mathcal{G}$ **(b)** Insertion of arc-variable and obtaining $\mathcal{G}'$ **(c)** Intervention on the arc-variable

Figure 11: Intervention on a causal mechanism and construction of extended graph $\mathcal{G}'$

To represent the intervention on the mechanism $X \to Y$ as an intervention on a variable, we propose to introduce the variable X^Y, mirroring X within the equation for Y, thus transforming the *FCM* Ψ into Ψ'.

Definition 5.2 *(Modified FCM isolating the causal mechanism from X to Y)*

Let Ψ be an FCM representing the relationships between the variables $\{V1 \cdots, V_N, X, Y\}$. To represent interventions on the causal mechanism $X \to Y$, we define an equivalent FCM Ψ' as a copy of the structural equations of Ψ, except for the insertion of X^Y as a virtual copy of X:

$$\Psi \begin{cases} V_1 = F_{V_1}(Pa(V_1), u_1) \\ \cdots \\ X = F_X(Pa(X), u_X) \\ Y = F_Y(Pa(Y), u_Y) \end{cases} \xrightarrow{becomes} \Psi' \begin{cases} V_1 = F_{V_1}(Pa(V_1), u_1) \\ \cdots \\ X = F_X(Pa(X), u_X) \\ X^Y = Id(X) \\ Y = F_Y(Pa(Y) \setminus \{X\}, X^Y, u_Y) \end{cases}$$

Intervening on the mechanism $X \to Y$ in Ψ is then equivalent to intervening on the variable X^Y in Ψ'.

This definition of an intervention on a causal mechanism can be transposed into the domain of causal graphs. From a graphical perspective, this intervention is equivalent to a two-step modification of the graphical model: (i) inserting a virtual copy X^Y of X between X and Y, and then (ii) performing an intervention $do(X^Y = x)$ on the new variable. Inserting variables in causal arcs transforms a graphical model $\mathcal{G}$ into $\mathcal{G}'$, where the arc $X \to Y$ from $\mathcal{G}$ can be replaced with $X \to X^Y \to Y$ in $\mathcal{G}'$, X^Y being a virtual copy of X.

It's worth remarking that the insertion of a variable between X and Y doesn't disrupt the graph's acyclicity. Thus, maintaining the link between *FCM* and Bayesian Networks as discussed in the paragraph following Definition 2.4.

Definition 5.3 *(Extended graph $\mathcal{G}'$ and Arc-variable X^Y) Let $\mathcal{G}$ be a causal graph representing a FCM Ψ. The extended graph $\mathcal{G}'$ is obtained from $\mathcal{G}$ by inserting a variable between X and Y and its FCM is the modified FCM Ψ', associated with the causal mechanism between X and Y.*

This newly introduced variable is identified as an arc-variable, also termed an $\mathcal{A}$-variable, labeled as X^Y. It functions as a virtual copy of X inserted between X and Y. By construction $\mathcal{G}'$ respects the following criterion:

- $\forall v \in \mathcal{V} \setminus \{Y\}$, $Pa_{\mathcal{G}}(v) = Pa_{\mathcal{G}'}(v)$ *where* $Pa_{\mathcal{G}}(v)$ *(resp.*$Pa_{\mathcal{G}'}(v)$*) is the set of parents of v in $\mathcal{G}$ (resp.$\mathcal{G}'$).*
- $Pa_{\mathcal{G}'}(X^Y) = \{X\}$
- $Pa_{\mathcal{G}'}(Y) = \{X^Y\} \cup Pa_{\mathcal{G}}(Y) \setminus \{X\}$

Having observed the translation of the definition into graphical terms, we hereby present this definition applied to the probability within the new graph $\mathcal{G}'$.

Definition 5.4 *(Probabilities of the extended graphical model)*

Let $\mathcal{G}$ be a causal graphical model representing the set of variable $\mathcal{V}$ and $\mathbb{P}_{\mathcal{G}}$ its joint distribution. Let $\mathcal{G}'$ the causal graphical model obtained by inserting an $\mathcal{A}$-variable X^Y into the arc $X \to Y$ of $\mathcal{G}$, $\mathcal{V}'$ the variables of $\mathcal{G}'$ and $\mathbb{P}_{\mathcal{G}'}$ its joint distribution. By construction, we then have :

$$\textit{homogeneity} \begin{cases} \forall v \in \mathcal{V} \setminus \{Y\},\ Pa_{\mathcal{G}}(v)) = Pa_{\mathcal{G}'}(v) \\ \forall v \in \mathcal{V} \setminus \{Y\},\ \mathbb{P}_{\mathcal{G}}(v|Pa_{\mathcal{G}}(v)) = \mathbb{P}_{\mathcal{G}'}(v|Pa_{\mathcal{G}'}(v)) \end{cases}$$

$$\textit{copy insertion} \begin{cases} \Omega(X) = \Omega(X^Y) \\ Pa_{\mathcal{G}}(Y) \setminus \{X\} = Pa_{\mathcal{G}'}(Y) \setminus \{X^Y\} \\ \forall (x, x') \in \Omega(X)^2, \mathbb{P}_{\mathcal{G}'}(X^Y = x'|X = x) = [\![x' = x]\!] \end{cases}$$

$$\textit{copy replacement} \begin{cases} \mathbb{P}_{\mathcal{G}'}(Y|X^Y = x) = \mathbb{P}_{\mathcal{G}}(Y|X = x) \\ \forall x \in \Omega(X), \mathbb{P}_{\mathcal{G}'}(Y|Pa_{\mathcal{G}'}(Y), X^Y = x) = \mathbb{P}_{\mathcal{G}}(Y|Pa_{\mathcal{G}}(Y), X = x) \end{cases}$$

where $\Omega(X)$ the range of X and $[\![x = x']\!]$ is the Iverson bracket notation:

$$[\![x = x']\!] = \begin{cases} 1 & \textit{if } x = x' \\ 0 & \textit{otherwise} \end{cases}$$

Inserting an $\mathcal{A}$-variable does not change the probabilistic model for the pre-existing variables, so the distributions and join probabilities of pre-existing variables are unchanged. Thus, we get the following rules to rewrite joint probabilities involving a mix of pre-existing and $\mathcal{A}$-variables.

Theorem 5.1 *(The joint distribution of $\mathcal{G}'$ comes from $\mathcal{G}$)*
Let $\mathcal{G}$ be a causal graphical model representing the set of variable $\mathcal{V}$. Let $\mathcal{G}'$ the causal graphical model obtained by inserting the $\mathcal{A}$-variable X^Y into the arc $X \to Y$ of $\mathcal{G}$, and $\mathcal{V}'$ the variables of $\mathcal{G}'$. We have:

$$\forall (x, x') \in \Omega(X),\ \mathbb{P}_{\mathcal{G}'}(\mathcal{V}' \setminus \{X^Y, X\}, X^Y{=}x', X{=}x) = [\![x = x']\!]\mathbb{P}_{\mathcal{G}}(\mathcal{V} \setminus \{X\}, X{=}x)$$

Proof: of Theorem 5.1

$$\begin{aligned}
\mathbb{P}_{\mathcal{G}}(\mathcal{V}) &= \prod_{v \in \mathcal{V}} \mathbb{P}_{\mathcal{G}}(v|Pa_{\mathcal{G}}(v)) = \prod_{v \in \mathcal{V}\setminus\{Y\}} \mathbb{P}_{\mathcal{G}}(v|Pa_{\mathcal{G}}(v)) \times \mathbb{P}_{\mathcal{G}}(Y|Pa_{\mathcal{G}}(Y)) \\
&= \prod_{v \in \mathcal{V}\setminus\{Y\}} \mathbb{P}_{\mathcal{G}'}(v|Pa_{\mathcal{G}'}(v)) \times \mathbb{P}_{\mathcal{G}}(Y|Pa_{\mathcal{G}}(Y) \setminus X, X) \\
&= \prod_{v \in \mathcal{V}\setminus\{Y\}} \mathbb{P}_{\mathcal{G}'}(v|Pa_{\mathcal{G}'}(v)) \times \mathbb{P}_{\mathcal{G}'}(Y|Pa_{\mathcal{G}'}(Y) \setminus X^Y, X^Y = X) \\
&= \prod_{v \in \mathcal{V}\setminus\{Y\}} \mathbb{P}_{\mathcal{G}'}(v|Pa_{\mathcal{G}'}(v)) \times \mathbb{P}_{\mathcal{G}'}(Y|Pa_{\mathcal{G}'}(Y)) = \prod_{v \in \mathcal{V}} \mathbb{P}_{\mathcal{G}'}(v|Pa_{\mathcal{G}'}(v))
\end{aligned}$$

$$\begin{aligned}
&\forall (x, x') \in \Omega(X), \\
&\mathbb{P}_{\mathcal{G}'}(\mathcal{V}') = \prod_{v \in \mathcal{V}'} \mathbb{P}_{\mathcal{G}'}(v|Pa_{\mathcal{G}'}(v)) = \prod_{v \in \mathcal{V}'\setminus\{X^Y\}} \mathbb{P}_{\mathcal{G}'}(v|Pa_{\mathcal{G}'}(v)) \times \mathbb{P}_{\mathcal{G}'}(X^Y = x'|X = x) \\
&\quad = \prod_{v \in \mathcal{V}} \mathbb{P}_{\mathcal{G}'}(v|Pa_{\mathcal{G}'}(v)) \times \mathbb{P}_{\mathcal{G}'}(X^Y{=}x'|X = x) = \mathbb{P}_{\mathcal{G}}(\mathcal{V}) \times \mathbb{P}_{\mathcal{G}'}(X^Y{=}x'|X{=}x) \\
&\quad = \mathbb{P}_{\mathcal{G}}(\mathcal{V}) \times [\![x' = x]\!]
\end{aligned}$$

□

Proposition 5.1 *(Rewriting rules for $\mathcal{A}$-variables)*
Let $\mathcal{G}$ be a causal graphical model representing the set of variable $\mathcal{V}$. Let $\mathcal{G}'$ the graphical model obtained by inserting $\mathcal{A}$-variables into $\mathcal{G}$, and $\mathcal{V}'$ the variables of $\mathcal{G}'$.

(a) $\forall S \subset \mathcal{V},\ \mathbb{P}_{\mathcal{G}'}(S) = \mathbb{P}_{\mathcal{G}}(S)$

(b) $\forall S \subset \mathcal{V}' \setminus \{X, X^Y\},\ \mathbb{P}_{\mathcal{G}'}(S, X = x, X^Y = x') = [\![x' = x]\!]\mathbb{P}_{\mathcal{G}'}(S, X = x)$

(c) $\forall S \subset \mathcal{V}' \setminus \{X, X^Y\},\ \mathbb{P}_{\mathcal{G}'}(S, X^Y = x) = \mathbb{P}_{\mathcal{G}'}(S, X = x)$

Proof: of Prop.5.1

(a) $\forall S \subset \mathcal{V},\ \mathbb{P}_{\mathcal{G}'}(S) = \mathbb{P}_{\mathcal{G}}(S)$:

$$\mathbb{P}_{\mathcal{G}}(S) = \sum_{v \notin S} \mathbb{P}_{\mathcal{G}}(\mathcal{V}) = \sum_{v \notin S} \prod_{v \in \mathcal{V}} \mathbb{P}_{\mathcal{G}}(v|Pa_{\mathcal{G}}(v))$$

From Theorem 5.1:

$$= \sum_{v \notin S} \prod_{v \in \mathcal{V}} \mathbb{P}_{\mathcal{G}'}(v|Pa_{\mathcal{G}'}(v)) = \sum_{v \notin S} \mathbb{P}_{\mathcal{G}'}(\mathcal{V}) = \mathbb{P}_{\mathcal{G}'}(S)$$

Thus : $\forall S \subset \mathcal{V},\ \mathbb{P}_{\mathcal{G}'}(S) = \mathbb{P}_{\mathcal{G}}(S)$

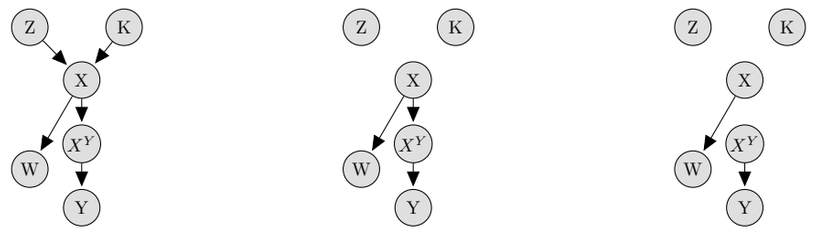
(b) $\forall S{\subset}\mathcal{V}'\setminus\{X, X^Y\}, \forall (x, x'){\in}\Omega(X),\ \mathbb{P}_{\mathcal{G}'}(S, X{=}x, X^Y{=}x'){=}[\![x' = x]\!]\mathbb{P}_{\mathcal{G}'}(S, X{=}x)$

$$\mathbb{P}_{\mathcal{G}'}(S, X^Y{=}x', X{=}x){=}\mathbb{P}_{\mathcal{G}'}(S|X^Y{=}x', X{=}x)\mathbb{P}_{\mathcal{G}'}(X^Y{=}x'|X{=}x)\mathbb{P}_{\mathcal{G}'}(X{=}x)$$

over and above $\mathbb{P}_{\mathcal{G}'}(X^Y{=}x'|X{=}x){=}[\![x'{=}x]\!]$, hence in the following we have $x{=}x'$. Thus:

$$\begin{aligned}
\mathbb{P}_{\mathcal{G}'}(S, X^Y = x', X = x) &= [\![x' = x]\!]\mathbb{P}_{\mathcal{G}'}(S|X^Y = x', X = x)\mathbb{P}_{\mathcal{G}'}(X = x) \\
&= [\![x' = x]\!]\mathbb{P}_{\mathcal{G}'}(S|X^Y = x, X = x)\mathbb{P}_{\mathcal{G}'}(X = x) \\
&= [\![x' = x]\!]\mathbb{P}_{\mathcal{G}'}(S|X = x)\mathbb{P}_{\mathcal{G}'}(X = x) \\
&= [\![x' = x]\!]\mathbb{P}_{\mathcal{G}'}(S, X = x)
\end{aligned}$$

(c) $\forall S \subset \mathcal{V}' \setminus \{X, X^Y\},\ \mathbb{P}_{\mathcal{G}'}(S, X^Y = x') = \mathbb{P}_{\mathcal{G}'}(S, X = x')$



(a) Pre-intervention **(b)** $do(X = x)$ **(c)** $do(X = x, X^Y = x)$

Figure 12: Insertion of an intervention on an $\mathcal{A}$-variable.

$$\begin{aligned}\mathbb{P}_{\mathcal{G}'}(S, X^Y = x') &= \sum_x \mathbb{P}_{\mathcal{G}'}(S, X^Y = x', X = x) \\ &= \sum_x [\![x' = x]\!] \mathbb{P}_{\mathcal{G}'}(S, X = x)\end{aligned}$$

Only the terms with x' remain:

$$\mathbb{P}_{\mathcal{G}'}(S, X^Y = x') = \mathbb{P}_{\mathcal{G}'}(S, X = x')$$

□

Intervening on the causal mechanism we defined as modifying the direct effect of X on Y is the same as intervening on the $\mathcal{A}$-variable X^Y in the extended graph $\mathcal{G}'$. This setup enables us to employ do-calculus and Property 5.1 to identify and estimate interventions on $\mathcal{A}$-variables with observable data.

Usual formulas about causal effects can still be used in $\mathcal{G}'$. Let $\mathcal{V}$ be the set of variables of $\mathcal{G}$. This set is a subset of the variables of $\mathcal{G}'$, and values are transported from $\mathcal{G}$ to $\mathcal{G}'$. Since the arcs inserted in $\mathcal{G}'$ have the same direction as the arcs they replace, an information path involving variables in $\mathcal{V}$ is active (resp. blocked) in $\mathcal{G}'$ if and only if it is also active (resp. blocked) in $\mathcal{G}$ when conditioning on any subset of $\mathcal{V}$. Thus, do-calculus rules applicable in $\mathcal{G}$ can also be applied in $\mathcal{G}'$, and the total causal effect of a $\mathcal{G}$ variable can be estimated in $\mathcal{G}'$ as in previous sections.

Lastly, we can derive from the identity $X^Y = Id(X)$ additional properties about interventions on $\mathcal{A}$-variables.

Proposition 5.2 *(Insertion of interventions on $\mathcal{A}$-variables)*

Let $\mathcal{G}$ be a causal graphical model, $\mathcal{G}'$ the causal graphical model obtained by inserting $\mathcal{A}$-variables into $\mathcal{G}$, X and Y variables copied from to $\mathcal{G}$ to $\mathcal{G}'$, and X^Y the $\mathcal{A}$-variable from X to Y in $\mathcal{G}'$. The intervention $do(X = x)$ in $\mathcal{G}'$ is equivalent to the intervention $do(X = x, X^Y = x)$.

Proof: Applying the $do(.)$ operator to X is equivalent to removing the arrows directed towards X, while preserving the arrows going away from X, and forcing the value of X to x [28]. Since $X^Y = Id(X)$, this intervention forces the value X^Y to x, which can only propagate from X^Y to its single child Y. It follows that the intervention on X (Fig. 12b) is equivalent to an intervention on both X and X^Y (Fig. 12c). □

Proposition 5.3 *(Equivalence between an intervention on a variable and interventions on its outgoing arcs)*

Let $\mathcal{G}$ be a causal graphical model, $\mathcal{G}'$ the causal graphical model obtained by inserting $\mathcal{A}$-variables into $\mathcal{G}$. Let Y, X, $X^\star$, K and W five disjoint sets of variables of $\mathcal{G}'$, where X is a single variable copied from $\mathcal{G}$ and $X^\star$ is the set of $\mathcal{A}$-variables inserted between X and each of its children in $\mathcal{G}$.

$$\mathbb{P}_{\mathcal{G}'}(y|do(x), do(k), w) = \mathbb{P}_{\mathcal{G}'}(y|do(X^\star = x), do(k), w)$$

where $do(X^\star = x)$ represents the $do(.)$ intervention on all variables of $X^\star$, forcing them to x.

Proof: of 5.3 From repeated applications of property 5.2, the intervention on X is equivalent to an intervention on X and $X^\star$:

$$\mathbb{P}_{\mathcal{G}'}(y|do(x), do(k), w) = \mathbb{P}_{\mathcal{G}'}(y|do(x), do(X^\star = x), do(k), w)$$

That we can reorder as:

$$\mathbb{P}_{\mathcal{G}'}(y|do(X = x), do(k), w) = \mathbb{P}_{\mathcal{G}'}(y|do(X^\star = x, k), do(x), w)$$

After removing from $\mathcal{G}'$ the arrows into X and the arrows into each variable in $X^\star$, X becomes disconnected, and thus $(Y \perp\!\!\!\perp X | X^\star, K, W)_{\mathcal{G}'_{\overline{X^\star X}}}$ (where $\mathcal{G}'_{\overline{X^\star, X}}$ is the causal graph obtained by removing all arrows pointing to nodes in $X^\star, X$). Removing further arrows would not invalidate this independence, and variables in X do not have any descendant once all arrows into $X^\star$ have been deleted, so the condition to apply rule 3 of do-calculus is satisfied, and:

$$\mathbb{P}_{\mathcal{G}'}(y|do(x), do(k), w) = \mathbb{P}_{\mathcal{G}'}(y|do(X^\star = x, k), w)$$

□

(a) Causal Bayesian Model. **(b)** Model with $\mathcal{A}$-variable.

Figure 13: Graphical models before and after adding $\mathcal{A}$-variable.

5.2.1 Identification and estimation of direct effects

To analyze the direct effect represented by the causal arc $X \to Y$, we can assess the outcome of an intervention $do(X^Y = x)$ in $\mathcal{G}'$. Identification of this intervention can rely on do-calculus, with the caveat that $\mathbb{P}_{\mathcal{G}'}(...|X^Y = x, X = x', ...)$ is only defined for $x = x'$, due to $\mathbb{P}_{\mathcal{G}'}(X^Y = x, X = x') = 0$ for $x \neq x'$.

Proposition 5.4 *(Controlled direct effect)*
We can infer that for all mediator M (child of X and parent of Y):

$$\mathbb{P}_{\mathcal{G}}(Y|do(X = x, M = m)) = \mathbb{P}_{\mathcal{G}'}(Y|do(X^Y = x, M = m))$$

Proof: of 5.4
From proposition 5.2,

$$\mathbb{P}_{\mathcal{G}'}(y|do(x), do(m)) = \mathbb{P}_{\mathcal{G}'}(y|do(x), do(X^Y = x, m))$$

After removing arrows into X, X^Y and M, X is disconnected, so the condition $(Y \perp\!\!\!\perp X|X^Y, M)_{\mathcal{G}'_{\overline{X^Y, MX}}}$ of rule 3 of do-calculus is satisfied and we can delete the intervention on X:

$$\mathbb{P}_{\mathcal{G}'}(y|do(x), do(m)) = \mathbb{P}_{\mathcal{G}'}(y|do(X^Y = x), do(m))$$

The same rules of do-calculus apply in $P_{\mathcal{G}}$ and $P_{\mathcal{G}'}$ for variables common to the two graphs, and variables in $P_{\mathcal{G}'}$ are identical copies of those in $P_{\mathcal{G}}$, so any rewriting of $\mathbb{P}_{\mathcal{G}}(y|do(x), do(m))$ is valid in $P_{\mathcal{G}'}$, and:

$$\mathbb{P}_{\mathcal{G}'}(y|do(X^M = x), do(m)) = \mathbb{P}_{\mathcal{G}}(y|do(x), do(m))$$

□

This property allows us to propose an alternate definition of the Controlled Direct Effect (*CDE*) in the modified graphical model $\mathcal{G}'$ (Fig.13b):

$$CDE_{\mathcal{G}'}(u, u', d) = \mathbb{P}_{\mathcal{G}'}(R|do(U^R = u, D = d)) - \mathbb{P}_{\mathcal{G}'}(R|do(U^R = u', D = d))$$

From proposition 5.4:

$$CDE_{\mathcal{G}'}(u, u', d) = \mathbb{P}_{\mathcal{G}}(R|do(U = u, D = d)) - \mathbb{P}_{\mathcal{G}}(R|do(U = u', D = d))$$

We retrieve the formula of the classic definition of the *CDE*.

5.2.2 Exploring new causal questions

Now that we can represent an intervention on a causal mechanism using the $do(.)$ operator, we can leverage the large corpus about do-calculus to address further questions about direct causal effects.

For instance, we might want to quantify the effect of a hypothetical intervention that would increase *Usage* without impacting *Discount*. A real-world equivalent could be a spot intervention that observes current usage, grants any discount accordingly, and then triggers an increment in usage. Let's define the direct effect of a hypothetical intervention that specifically alters the causal mechanism from U to R, from an observed value $U^R = u$ to the modified value $U^R = u'$, as:

Definition 5.5 *(Direct Effect Through $\mathcal{A}$-variables)*

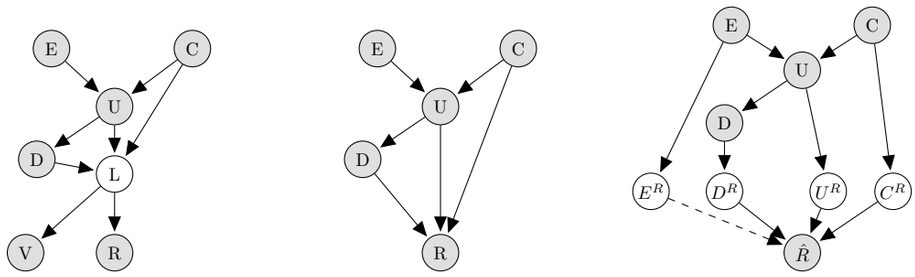
$$DE(u, u') = \mathbb{P}(R = 1|do(U^R = u'), U = u) - \mathbb{P}(R = 1|U = u)$$

Let's assume that Usage is represented by a finite set of integer values $\{0, 1, ..., l\}$. By computing the direct effect of increasing Usage by 1 for all its discrete values and by averaging the results over the observed distribution of Usage, we can estimate the direct causal effect of growing usage without modifying the budget allocated to discounts as:

$$\sum_{0 \leq u < l} DE(u, u+1)\mathbb{P}(u) + \mathbb{P}(R = 1|U = l)\mathbb{P}(U = l) - \mathbb{P}(R = 1)$$

Do-calculus computations in PyAgrum identify $DE(u, u')$ with:

$$DE(u, u') = \frac{\sum_{c,d,e} \mathbb{P}(R = 1|c, d, u')\mathbb{P}(d|u)\mathbb{P}(u|c, e)\mathbb{P}(c)\mathbb{P}(e)}{\mathbb{P}(u)} - \mathbb{P}(R = 1|u)$$



(a) Ground Truth **(b)** Causal Bayesian Model **(c)** Predictive Model

Figure 14: Ground Truth, Causal and Predictive Models

The formula from PyAgrum relies on the Conditional Probability Tables associated with nodes in a graphical model. Since we intend to use a different technique for the estimate, we can use Markov factorization rules to simplify the formula:

$$DE(u,u') = \sum_{c,d} \mathbb{P}(R=1|c,d,u')\mathbb{P}(c,d|u) - \mathbb{P}(R=1|U=u)$$

This formula can indeed be estimated by training a predictive model f of R on $\{C,D,U\}$ and performing a Monte-Carlo integration of f on a sample population $\mathcal{P}_{U=u}$ filtered on $U=u$:

$$\hat{DE}(u,u') = \frac{1}{|\mathcal{P}_{U=u}|} \sum_{\mathcal{P}_{U=u}} f(C,D,U=u') - \mathbb{E}_{\mathcal{P}_{U=u}}(R=1)$$

5.3 XAI interpreted as a quantification of direct causal effects

5.3.1 Example

Let's consider a data generation process represented by the ground truth model in Fig. 14a. This is the model we used throughout the article, with each variable defined by its first letter. In this model, L is latent, and all other variables can be observed.

Let $\mathcal{G}$ be the graphical causal model relating R and all its direct or indirect causes (Fig. 14b). We assumed that domain expertise allowed us to properly exclude V from the set of causal ancestors of R ($An_{\mathcal{G}}(R) = \{C,D,E,U\}$), and that R has no unknown confounder.

The Markov boundary of R in $\mathcal{G}$ reduces to the set $Pa_{\mathcal{G}}(R) = \{C,D,U\}$ of direct parents of R in $\mathcal{G}$.

$$\mathbb{P}_{\mathcal{G}}(R|An_{\mathcal{G}}(R)) = \mathbb{P}_{\mathcal{G}}(R|Pa_{\mathcal{G}}(R)) \tag{14}$$

Finally, let's consider a model $f(.)$ trained on $An_{\mathcal{G}}(R)$ to produce predictions about R. The data extraction process used to feed the learning algorithm with training observations is represented in Fig.14c, with X^R=$Id(X)$.

A model learned according to the principles of the Statistical Learning Theory [41] is expected to produce minimal generalization errors on a population sampled from the same data generation process:

$$f(An_{\mathcal{G}}(R)) \approx \mathbb{P}_{\mathcal{G}}(R=1|Pa_{\mathcal{G}}(R))$$

The model Figure 14c is equivalent to a model $\mathcal{G}'$ with $\mathcal{A}$-variables between Y and its parents. A reasonable assumption when training a predictive model is that any known ancestor of R might have a small direct effect on it. From this assumption, we first insert the direct arc $E \to R$ before inserting E^R.

5.3.2 Partial Dependence Plots

Let Y be a target variable, A the set of its causal ancestors in the graph $\mathcal{G}$, $\mathcal{G}'$ the graph obtained by inserting a set M of $\mathcal{A}$-variables between A and Y.

We note a subset $S \subset M$, and $\bar{S} = M \setminus S$. The effect $\mathbb{P}_{\mathcal{G}'}(Y|do(x_S))$ of an intervention on S variables can be identified using the same do-calculus steps involved in the backdoor theorem for an intervention on a single variable. Using marginalization and the chain rule for joint probabilities:

$$\mathbb{P}_{\mathcal{G}'}(Y|do(x_S)) = \sum_{x_{\bar{S}}} \mathbb{P}_{\mathcal{G}'}(Y|do(x_S),x_{\bar{S}})\mathbb{P}_{\mathcal{G}'}(x_{\bar{S}}|do(x_S)) \tag{15}$$

By definition, each $\mathcal{A}$-variable in M has the target as an only child. Furthermore, all the parents of the target are present in M so $\bar{S}=M\setminus S$ blocks all backdoor paths from S into Y, and $(Y \perp\!\!\!\perp S|\bar{S})_{\mathcal{G}'_{\underline{S}}}$. By application of rule 2 of do-calculus, we obtain:

$$\mathbb{P}_{\mathcal{G}'}(Y|do(x_S),x_{\bar{S}}) = \mathbb{P}_{\mathcal{G}'}(Y|x_S,x_{\bar{S}}) \tag{16}$$

After cutting all arrows into $\bar{S}$, the only paths connecting S and $\bar{S}$ collide on Y, and $(\bar{S} \perp\!\!\!\perp S)_{\mathcal{G}'_{\bar{S}}}$. By application of rule 3 of do-calculus:

$$\mathbb{P}_{\mathcal{G}'}(x_{\bar{S}}|do(x_S)) = \mathbb{P}_{\mathcal{G}'}(x_{\bar{S}}) \tag{17}$$

From Equations 15, 16 and 17:

$$\mathbb{P}_{\mathcal{G}'}(Y|do(x_S)) = \sum_{x_{\bar{S}}} \mathbb{P}_{\mathcal{G}'}(Y|x_S, x_{\bar{S}})\mathbb{P}_{\mathcal{G}'}(x_{\bar{S}})$$

Since X_S and $X_{\bar{S}}$ are exact copies of observed variables, an intervention on a set of direct causal arcs, represented as a $do(.)$ operation on $\mathcal{A}$-variables, is thus identified as observable probabilities, that can be estimated by a Monte-Carlo integration over a sample population $\mathcal{B}$:

$$\mathbb{P}_{\mathcal{G}'}(Y = 1|do(X_S = x_S)) \approx \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} f(x_S, X_{\bar{S}}^{(i)})$$

We recognize Eq.3 involved in calculating the *PDP* of S.

Hence, estimating the direct effects of a cause set S on a target Y involves training a predictive model for Y using all its causal ancestors, and so excluding other known variables. Then, the *PDP* can be used to determine the effects of S.

In our example, to estimate direct causal effects on R, we propose to train a predictive model of R on the feature selection $\{E, C, D, U\}$, and then apply the *PDP* formula to each input feature.

5.3.3 Marginal Shapley values

The characteristic function of marginal Shapley values is v^{marg} defined for a set S of variables as: $v^{marg}(S) = \mathbb{E}\big[f(x_S, X_{\bar{S}})\big]$.

Where S and $\bar{S}$ have the same definition as the previous paragraph 5.3.2.

v^{marg} is typically estimated over a sample background population $\mathcal{B}$ as [5]:

$$v^{marg}(S) = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} f(x_S, X_{\bar{S}}^{(i)})$$

Again, we recognize the equation 3 of the *PDP* of S: each calculation of the characteristic function v^{marg} on a coalition S of direct causes of Y amounts to the calculation of interventions on the causal arcs from S to Y.

Thus, provided that a predictive model has been trained on the set of direct (or ancestral) causes of the target, at the exclusion of other variables, marginal Shapley values give a quantification of all direct effects on Y under the constraints of efficiency, linearity, symmetry and nullity. For our example, the Shapley values in Fig.5b estimate the direct effects of four input features on *Renewal*, whereas the Shapley values in Fig.5a do not support straightforward causal interpretations about the data generation process (aside from *Economy* being outside the Markov boundary of the target).

Calculating marginal Shapley values involves considering every conceivable combination of variable interventions. However, not all interventions are practical or feasible in real-world settings. In our example: altering a customer's profile might be unfeasible while encouraging usage and applying discounts can be executed. Hence, employing the *PDP* technique on a limited variable set is more pertinent for estimating actionable causal effects.

In this section, we showed how incorporating a copy variable allowed us to effectively isolate and estimate the direct effect. This technique was then applied within our case study, where knowledge of the causal structure allowed for identifying the effect. Using the methodology detailed earlier, we estimated this direct effect.

6 Conclusion

This paper presents a novel XAI approach for better quantifying causal effects from observational data and a graphical method for gaining insights into direct effects.

Bluntly applying XAI tools to classifiers trained on all known features without considering causality can lead to flawed interpretations. To address this issue, we propose a new framework that begins with a training population, a causal graph, and a specific query about a causal effect. Employing tools of causal inference, we identify the causal impact in terms of observable probabilities. Subsequently, we train a machine learning model to estimate these probabilities. Then, an interpretability method quantifies the requested causal effect from the model predictions. Additionally, to assess a direct effect of X on Y, we introduce $\mathcal{A}$-variables, effectively isolating the causal mechanism to perform an intervention.

In the XAI community, there is a debate between adhering to the model's behavior (*true-to-the-model*) or staying faithful to the data (*true-to-the-data*) [4]. Relaxing the constraint of a pre-existing model is a step towards enabling XAI to be faithful to both the model and the data for a variable of interest. However, addressing multiple causal queries might involve training several predictive models.

Our approach uses causality to build and analyze predictive models. Despite its practical value, it faces some challenges, particularly in the task of learning a causal graph when latent variables or multiple parents for specific nodes are involved. Various methods can at least yield a partially directed causal graph (PDAG). As a possible next step, we could explore how such a partial causal graph could be used in our approach.

Finally, we emphasize that our proposal is not intended to replace randomized controlled trials or carefully designed studies on natural experiments. Our purpose is instead to improve the insights that can be gained from observational data.

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values, 2020.

[2] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[3] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models, 2021.

[4] Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data?, 2020.

[5] Hugh Chen, Ian C. Covert, Scott M. Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions, 2022.

[6] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, KDD'16, pages 785–794, New York, NY, USA, 2016. ACM.

[7] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[8] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[9] Ashley Deeks. The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7):1829–1850, 2019.

[10] Marek J. Druzdzel and Herbert A. Simon. Causality in bayesian belief networks. In David Heckerman and E. H. Mamdani, editors, *UAI '93: Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence, The Catholic University of America, Providence, Washington, DC, USA, July 9-11, 1993*, pages 3–11. Morgan Kaufmann, 1993.

[11] Gaspard Ducamp, Christophe Gonzales, and Pierre-Henri Wuillemin. aGrUM/pyAgrum : a Toolbox to Build Models and Algorithms for Probabilistic Graphical Models in Python. In *PGM'20*, volume 138 of *PMLR*, pages 609–612, Skørping, Denmark, September 2020.

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

[13] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. In *Advances in Neural Information Processing Systems*, volume 33, pages 1229–1239. Curran Associates, Inc., 2020.

[14] Shunkai Fu and Michel C Desmarais. Markov blanket based feature selection: a review of past decade. In *Proceedings of the world congress on engineering*, volume 1, pages 321–328. Newswood Ltd. Hong Kong, China, 2010.

[15] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

[16] 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*, 2015.

[17] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, Jun. 2019. doi: 10.1609/aimag.v40i2.2850.

[18] Mahdi Hadj Ali, Yann Le Biannic, and Pierre-Henri Wuillemin. Interpreting predictive models through causality: A query-driven methodology. In *The International FLAIRS Conference Proceedings*, volume 36, 2023.

[19] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models, 2020.

[20] Paul W. Holland. Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18:449–484, 1988.

[21] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 2013.

[22] Dominik Janzing, Lenon Minorics, and Patrick Bloebaum. Feature relevance quantification in explainable ai: A causal problem. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2907–2916. PMLR, 26–28 Aug 2020.

[23] Max Kuhn and Kjell Johnson. *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC, 2019.

[24] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR, 2020.

[25] Verny Louis, Nadir Sella, Séverine Affeldt, Param Priya Singh, and Hervé Isambert. Learning causal networks with latent variables from multivariate information in genomic data. *Public Library of Science Computational Biology*, 2017.

[26] Scott M Lundberg and Su-In Lee. Shap. github.com/slundberg/shap, 2018.

[27] Judea Pearl. Markov networks. In *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, chapter 3, page 97. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[28] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[29] Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[30] Judea Pearl. The do-calculus revisited. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, page 3–11, Virginia, 2012. AUAI Press.

[31] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

[32] Judea Pearl and Thomas Verma. A theory of inferred causation. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 441–452, 1991.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

[34] Thilo Rieg, Janek Frick, Hermann Baumgartl, and Ricardo Buettner. Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms. *PLOS-ONE*, 2020.

[35] Lloyd S Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, 1953.

[36] Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 499–506, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[37] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[38] Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *Journal Of Machine Learning Research*, 2010.

[39] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278, 13–18 Jul 2020.

[40] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, 2021. doi: 10.1109/TNNLS.2020.3027314.

[41] V. Vapnik. *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer New York, 1999. ISBN 9780387987804.

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

[43] Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 721–729. PMLR, 13–15 Apr 2021.

[44] H. P. Young. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72, Jun 1985. ISSN 1432-1270.

[45] Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of business and economic statistics*, 2019, 2019. ISSN 0735-0015.