Interprétation causale des modèles prédictifs : paradoxes et solution

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

Interprétation causale des modèles prédictifs : paradoxes et solution

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
mahdi.hadj.ali@sap.com, yann.le.biannic@sap.com , pierre-henri.wuillemin@lip6.fr

Abstract

Les modèles de Machine Learning ont été largement adoptés ces dernières années en raison de leur efficacité et de leur polyvalence dans de nombreux domaines. Cependant, la complexité des modèles prédictifs a conduit à un manque d'interprétabilité notamment pour la prise de décision automatique. Des travaux récents ont amélioré l'interprétabilité générale en estimant les contributions des variables d'entrée à la prédiction d'un modèle pré-entraîné. Malgré ces progrès, les utilisateurs sont toujours à la recherche d'informations causales sur les mécanismes sous-jacents de génération des données. À cette fin, certains travaux ont tenté d'intégrer des connaissances causales dans l'interprétabilité, car les techniques non causales peuvent conduire à des explications paradoxales. Ces efforts ont permis de répondre à diverses questions, mais le fait de s'appuyer sur un seul modèle pré-établi peut entraîner des problèmes de quantification. Dans cet article, nous soutenons que chaque requête causale nécessite un raisonnement adéquat ; par conséquent, un modèle prédictif unique n'est pas adapté à toutes les questions. Au lieu de cela, nous proposons un nouveau cadre qui donne la priorité à la requête d'intérêt et du quel dérive ensuite une méthodologie axée sur la requête en fonction de la structure du modèle causal. Il en résulte un modèle prédictif sur mesure adapté à la requête et une technique d'interprétabilité adaptée. Plus précisément, elle fournit une estimation numérique des effets causaux, ce qui permet d'apporter des réponses précises aux questions d'interprétabilité lorsque la structure causale est connue. Ce papier reprend et traduit des contributions présentées à FLAIRS'23.

Introduction

Les modèle prédictifs récents sont de plus en plus sophistiquées et améliorent généralement la précision de prédiction, mais au prix d'une plus grande difficulté d'interprétation. De plus, l'interprétabilité de ces modèles est une question sensible dans de nombreux domaines (Burkart and Huber 2021). En effet, l'utilisation de modèles dans le cadre de la prise de décision automatique nécessite une connaissance détaillée de leur comportement afin de pouvoir justifier la décision ; par exemple, dans le domaine médical de la prescription automatique, dans le domaine juridique, ou dans un contexte légal (Rieg et al. 2020). Les utilisateurs sont souvent intéressés par des informations causales sur les mécanismes sous-jacents de génération des données, que les modèles prédictifs ne fournissent généralement pas. Parmi les questions causales courantes, citons l'identification des causes et des effets, la prévision des effets des interventions et la réponse aux questions contrefactuelles. Si nous supposons que le modèle causal sous-jacent du processus de génération des données peut être représenté sous la forme d'un réseau bayésien causal (c'est-à-dire un réseau bayésien où les orientations ont une interprétation causale), la solution idéale consiste à utiliser le cadre causal et des outils spécialisés tels que le do-calculus pour répondre à ces questions. Cependant, l'obtention du modèle causal complet peut s'avérer difficile en raison de la multiplicité des parents de la cible ou de l'impossibilité d'interroger les variables latentes. Par conséquent, il se peut que nous devions nous appuyer sur des hypothèses concernant uniquement sa structure causale et sur des modèles prédictifs.

Plusieurs travaux traitent de la quantification des effets causaux (directs et/ou indirects) à partir d'un modèle prédictif, en supposant que l'on connaisse la structure causale, qui décrit les connexions entre les variables (Heskes et al. 2020; Wang, Wiens, and Lundberg 2021). Ces études suivent le même schéma que le domaine de l'explicabilité de l'intelligence artificielle (XAI), c'est-à-dire qu'elles partent d'un modèle prédictif, généralement formé à partir de toutes les variables connues, et tentent ensuite de quantifier la contribution de chaque variable. L'objectif de cet article est de montrer les avantages d'une approche alternative où le modèle prédictif n'est plus donné mais est conçu pour répondre à une requête causale spécifique.

Comme dans les travaux précédents, nous supposons une connaissance préalable de la structure causale, mais nous proposons de l'utiliser avant de construire, d'entraîner et d'analyser des modèles prédictifs pilotés par des requêtes à partir de données d'observation.

La première partie de cet article présente les techniques de XAI les plus récentes et quelques notions de causalité. Ensuite, nous décrivons notre approche et la configuration qui nous permettra de comparer les différentes approches sur un ensemble de données synthétiques. La quatrième section montre les limites de l'utilisation d'un modèle prédictif prédéfini, généralement entraîné sur toutes les variables

d'entrée. Dans les dernières sections, nous étudierons deux questions causales qui sont difficiles à traiter avec les approches actuelles mais qui peuvent être correctement traitées par notre proposition.

Modèles prédictifs, modèles causaux

Modèles prédictifs et explicabilité

Une tâche courante en apprentissage supervisé consiste à prédire la classe binaire Y d'une cible à partir d'un vecteur de variables $\mathbf{X} = \{X_1, \cdots, X_j, \cdots, X_M\}$. Un modèle prédictif est formé à partir d'une base de données d'observations sur la classe et les variables. Il est défini comme une fonction à valeur réelle f qui prend un vecteur de variables en entrée et renvoie une estimation de la probabilité de la classe cible : $f(\mathbf{X}) \simeq P(Y = 1|\mathbf{X})$.

Plusieurs outils ont été développés pour expliquer les prédictions faites par un modèle ML. Par exemple, le *Partial Dependence Plots* (PDP) propose d'examiner l'effet de la j-ième variable en étudiant la prédiction moyenne lorsque cette variable est perturbée (Friedman 2001). Les *Individual Conditional Expectation Plots* (ICE) reposent sur la même idée que les PDP mais correspondent à l'étude de la prédiction par f à partir d'un exemple donné lorsque la j-ième variable est modifiée (Goldstein et al. 2015). Ainsi, la moyenne de tous les ICEs correspond au PDP.

Une autre idée consiste à estimer un score d'importance pour chaque variable. Breiman propose d'échanger une variable avec du bruit et d'évaluer l'impact sur les prédictions.

Cet article fera souvent référence aux valeurs de Shapley (Shapley 1953) et à leur application au XAI. Les valeurs de Shapley sont une méthode pour répartir le gain entre les joueurs $\mathbf{X}$ dans un "jeu coopératif" (von Neumann and Morgenstern 1947). Dans ce cadre, une fonction de valeur v associe un nombre réel $v(S)$ à toute coalition $S \subseteq \mathbf{X}$. Pour transposer ce cadre à l'explicabilité, un parallèle est établi entre la prédiction d'un modèle et la fonction de valeur d'un jeu, ainsi qu'entre les variables d'entrée $\mathbf{X}$ et les joueurs qui collaborent pour gagner $f(\mathbf{X})$.

Les valeurs de Shapley sont ainsi devenues un moyen d'expliquer un modèle et se sont répandues dans le domaine du ML (Strumbelj and Kononenko 2010; Lundberg and Lee 2017). Plusieurs variantes (Sundararajan and Najmi 2020; Frye, Rowat, and Feige 2020; Heskes et al. 2020; Wang, Wiens, and Lundberg 2021; Kolpaczki, Bengs, and Hüllermeier 2023) ont été proposées. Parmi celles-ci, nous ferons principalement référence aux valeurs SHAP très répandues (Lundberg and Lee 2017).

L'explication fournie par les valeurs SHAP est une excellente base pour comprendre le comportement d'un modèle prédictif. Les valeurs SHAP offrent une explication indépendante du modèle et reposent sur des bases mathématiques solides. Toutefois, le problème de l'explicabilité réside souvent davantage dans la prescription que dans la prédiction. L'analyse prédictive vise à répondre à des questions telles que "Quelle est la valeur probable de Y si j'ai observé X ?" ou "Quels sont les poids des paramètres qui conduisent à la prédiction ?". D'autre part, l'analyse prescriptive répond à des questions telles que "Quelle intervention devrais-je faire pour améliorer Y ?" et "Quand et pourquoi devrais-je faire une telle intervention ?". Les valeurs SHAP quantifient les contributions des variables à la prédiction faite par un modèle et répondent donc aux besoins de l'analyse prédictive. Toutefois, il est facile de confondre les contributions des variables et l'effet d'une intervention. Ce dernier est nécessaire pour l'analyse prescriptive.

Modèles causaux et explicabilité

Une solution potentielle aux questions prescriptives consiste à se tourner vers le cadre et les outils causaux. D'un point de vue causal, on peut répondre à ces questions par l'effet causal d'une variable actionnable sur la cible. Une variable actionnable est une variable sur laquelle on peut agir dans le "monde réel", c'est-à-dire que l'on peut intervenir sur la variable et donc contrôler sa valeur. Le do-calcul (Pearl 2012) est une solution pour l'estimation des effets causaux.

Janzing et al. (2013) propose de quantifier la contribution causale d'une variable binaire A par son effet causal moyen (ACE) :

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

Ceci est similaire au concept d'élévation moyenne (Rubin 1974; Gutierrez and Gérardy 2017; Devriendt, Moldovan, and Verbeke 2018), qui est apprécié pour sa simplicité et facilite ainsi la prise de décision.

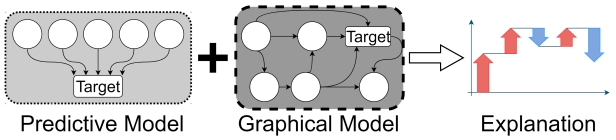

Figure 1: Approche classique du XAI.

Notre proposition de combiner les modèles causaux et prédictifs

Les méthodes de XAI visent généralement à expliquer les prédictions faites par un modèle préalablement entraîné. Certaines méthodes intègrent la causalité via un modèle graphique des relations causales sous-jacentes entre les variables (Frye, Rowat, and Feige 2020; Heskes et al. 2020). Toutefois, ces méthodes héritent du principe général du XAI selon lequel un seul modèle prédictif pré-entraîné est la principale source d'estimations pour répondre aux questions causales concernant de multiples variables, voir la figure 1.

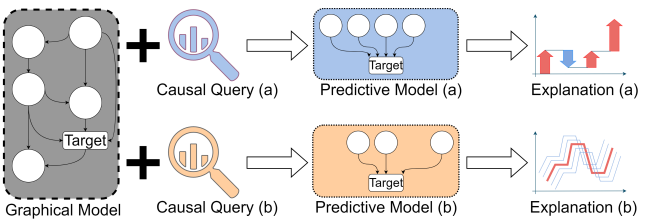

Figure 2: Approche proposée avec 2 requêtes distinctes.

Dans cet article, nous proposons une nouvelle méthodologie, illustrée par la figure 2, qui étend le cadre commun décrit ci-dessus. Notre approche comporte plusieurs phases. Tout d'abord, nous partons d'une population d'entraînement, d'un graphe causal et d'une requête causale spécifique. Ensuite, nous formons un modèle ML adéquat à la requête et au contexte causal. Enfin, nous utilisons une méthode d'interprétabilité adaptée à la requête et au contexte pour quantifier l'effet souhaité.

L'une des principales différences entre notre proposition et les méthodes précédentes est que nous ne partons pas d'un modèle pré-entraîné. L'argument principal est que les différentes questions causales ne peuvent pas être systématiquement résolues par un seul modèle prédictif général. La construction du modèle qui génère l'explication doit également tenir compte des contraintes imposées par le calcul causal.

Protocole expérimentale

Cet article examine la faisabilité de la quantification des effets causaux multiples à partir de données d'observation en utilisant des algorithmes d'apprentissage supervisé standard et des techniques d'interprétabilité. Dans la pratique, les modèles appris peuvent être biaisés ou altérés. Pour surmonter ces problèmes, nous proposons un réseau bayésien causal comme référence "ground truth".

Une base de données est générée à partir de ce modèle de référence. Les données sont ensuite utilisées comme base d'apprentissage pour les modèles prédictifs que nous essayons d'expliquer. L'utilisation d'un réseau bayésien causal comme "ground truth" nous permet de quantifier les effets causaux exacts des variables d'intérêt à l'aide de méthodes analytiques telles que do-calcul (Pearl 2000). Nous pouvons ainsi examiner les interprétations d'un modèle de classification et évaluer leur cohérence avec le modèle causal sous-jacent.

Pour illustrer notre propos, nous avons conçu un exemple synthétique avec pyAgrum (Ducamp, Gonzales, and Wuillemin 2020), une bibliothèque pour les modèles graphiques probabilistes. Pour faciliter le raisonnement, nous avons attribué une sémantique à cet exemple : la tâche de prédiction est de savoir si le client va renouveler son abonnement de téléphone portable. La prédiction est basée sur plusieurs variables :

- *Economy* (noté E) représente les conditions économiques, de l'expansion ou la contraction,
- Le profil du client (professionnel ou privé) est représenté par la variable *Customer Profil* (notée C),
- la consommation annuelle du service par le client est suivie par la variable *Usage* (notée U),
- une offre unique accordée au client est illustrée par *Discount* (noté D),
- la fidélité du client ne peut être observée directement et sera traitée comme une variable latente *Loyalty* (notée L),
- *Visits* (noté V) indique si le client a récemment visité le site web du fournisseur,

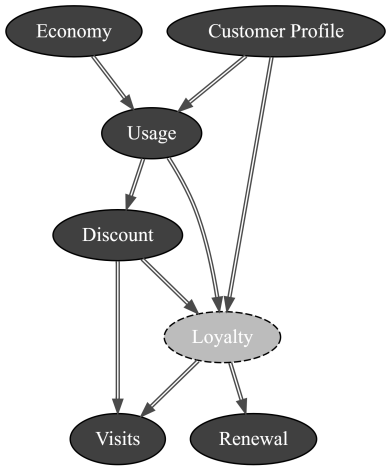

Figure 3: Le Réseau bayésien causal qui génère les données. *Loyalty* est latent.

- Finalement *Renewal* (noté R) informe sur le renouvellement de l'abonnement et sera la cible de la classification binaire.

Pour limiter la taille de l'espace des variables et former des modèles de classification précis, la plupart des variables sont binaires, à l'exception de *Usage*, qui peut prendre cinq valeurs distinctes. La figure 3 représente le réseau bayésien causal utilisé pour générer les données.

En pratique, deux explications intéressantes seraient l'effet de *Economy* et de *Discount*. Le modèle fictif a été conçu de manière à ce que l'accord d'une remise (*D=1*) ait un effet causal positif sur les renouvellements pour un profil de client (*C=0*) et aucun effet causal pour l'autre profil (*C=1*) :

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

et:

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

De même, *Economy* (*E=1*) a un effet total négatif sur *Renewal* lorsque *C=0* et aucun effet causal lorsque *C=1*.

L'objectif des sections suivantes est de montrer, dans différents contextes, comment l'interprétation causale des résultats XAI classiques peut être ambiguë (section) et comment notre proposition peut conduire à des estimations plus cohérentes des effets causaux à partir de prédiction spécifiques (section et). La mise en œuvre des exemples est fournie sous la forme d'un notebook sur GitHub[1].

Sensibilité à la sélection des variables

Dans toute analyse de données d'observation, il est bien connu que la sélection des variables a un impact significatif sur les modèles prédictifs. Cette section illustre comment cette sélection sans analyse causale peut conduire à des paradoxes (sous-section 1) ou à une quantification inutile/manquante de certains paramètres (sous-section 2).

[1]Les figures et les modèles peuvent être trouvés dans https://github.com/anonyme/query-driven-xai

Paradoxes de l'approche XAI classique

Pour former des modèles prédictifs à partir de populations, nous utilisons un algorithme de ML bien établi, XGBoost (Chen and Guestrin 2016). Nous entraînons deux modèles sur le même ensemble de données, mais en utilisant différents ensembles de variables d'entrée. Un premier modèle est formé sur toutes les variables connues (c'est-à-dire toutes les variables à l'exception de la cible et de la variable non observée *Loyalty*), et un second modèle est formé après le retrait de *Visits*. Pour ces deux modèles, nous utilisons la bibliothèque SHAP (Lundberg and Lee 2018) pour calculer les contributions des variables. Les résultats sont présentés dans les figures 4 et 5.

Figure 4: Représentation des contributions de SHAP, expliquant un modèle entraîné sur toutes les variables.

Figure 5: Représentation des contributions de SHAP, expliquant un modèle formé en excluant *Visits*.

Les deux graphiques représentent, comme indiqué dans la bibliothèque SHAP, les valeurs SHAP de chaque variable pour chaque échantillon. Le graphique trie les variables par la somme des valeurs SHAP sur tous les échantillons et utilise les valeurs SHAP pour montrer la distribution de l'impact de chaque variable sur la sortie du modèle. La couleur représente la valeur de la variable (rouge élevé, bleu faible).

La lecture de ces deux graphiques suggère que l'obtention d'une remise (points rouges de la variable *Discount*) contribue négativement aux prédictions dans le premier modèle Figure 4, alors qu'il a une contribution positive dans le second modèle Figure 5. Si les contributions étaient naïvement interprétées comme des effets causaux sur la cible, un analyste pourrait tirer des conclusions opposées des deux modèles. Dans cet exemple, nous observons que la méthode d'interprétation SHAP très répandue est sensible à la sélection des variables : elle peut fournir des indications contradictoires lorsqu'elle est appliquée à différents modèles formés à l'aide du même algorithme ML sur le même ensemble de données, mais sur des sélections de variables différentes.

Pouvoir prédictif versus effet causal

Plusieurs auteurs ont proposé d'intégrer la connaissance de la structure causale lors de l'interprétation d'un modèle prédictif. Cependant, la quantification des effets causaux peut nécessiter des informations qui ne peuvent être extraites du modèle. En effet, le modèle prédictif peut ne pas utiliser une variable qui a un effet causal indirect sur la cible. Cette situation se présente lorsque la variable est indépendante de la cible après conditionnement sur d'autres variables d'entrée.

Si nous supposons que SHAP représente correctement les contributions des variables d'entrée aux prédictions (Janzing, Minorics, and Bloebaum 2020), nous pouvons observer cette situation dans notre exemple synthétique. *Economy* a un effet causal indirect sur la cible : dans le modèle de génération de données, son effet causal moyen (ACE) est d'environ -2,8%. Cependant, l'effet causal de *Economy* passe par un médiateur (*Usage*) qui est une variable d'entrée du modèle prédictif. Ainsi, *Economy* n'apporte aucune information supplémentaire sur la cible par rapport à *Usage* et peut être ignoré par le modèle sans aucun impact sur la précision de la prédiction. En effet, nous observons à partir des valeurs SHAP extraites de nos deux modèles prédictifs que la contribution de *Economy* est proche de zéro.

D'autre part, une variable peut contribuer fortement à un modèle prédictif tout en n'étant ni une cause directe ou indirecte, ni une conséquence directe ou indirecte de la cible. Dans notre exemple synthétique, *Visits* n'est ni une cause ni une conséquence des renouvellements, mais l'existence d'une variable latente (*Loyalty*) implique que *Visits* n'est pas indépendante de la cible lorsqu'elle est conditionnée par toutes les variables connues, et donc que *Visits* apporte des informations supplémentaires sur la cible. En effet, les graphiques SHAP montrent que *Visits* est le meilleur prédicteur pour le modèle qui a accès à cette variable.

Quantification d'un effet total

Supposons que l'objectif soit d'évaluer l'effet d'une remise sur le renouvellement des abonnements. Dans cette section, nous montrons comment calculer exactement cet effet sous l'hypothèse d'un modèle causal complet, puis comment une application des outils XAI à partir de données d'observation guidée par une requête causale permet une approximation fiable de l'effet, même en présence de variables latentes.

Solution exacte à l'aide du do-calcul

Dans un cadre causal probabiliste, l'interrogation sur l'effet causal total est la quantification de la probabilité $P(Y|do(X))$. Dans un tel cadre, do-calcul fournit plusieurs techniques, telles que les ajustements Frontdoor ou Backdoor, pour calculer les effets causaux (Pearl 2000). En particulier, l'ajustement du Backdoor définit un ensemble de variables à prendre en compte pour pouvoir calculer l'effet recherché.

Définition *(critère du Backdoor) - Étant donné une paire ordonnée de variables (X, Y) dans un graphe acyclique dirigé G, un ensemble de variables Z satisfait le critère du Backdoor par rapport à (X, Y) :*

- *(i) si aucun noeud de Z n'est un descendant de X, et*
- *(ii) Z bloque tout chemin entre X et Y qui contient une flèche vers X.*

Si un ensemble de variables Z satisfait le critère du Backdoor relativement à (X, Y), alors l'effet causal de X sur Y est identifiable et est donné par la formule suivante:

Définition *(Ajustement du Backdoor) - Si Z satisfait le critère du Backdoor par rapport à (X, Y) :*

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

Appliqué à notre exemple (figure 3), la formule 1 permet de quantifier l'effet causal de *Discount* sur *Renewal* avec $\{Usage\}$ en tant qu'ensemble satisfaisant le critère du Backdoor.

Estimations à partir d'un échantillon de données

L'estimation de l'effet causal au moyen de l'ajustement du Backdoor dans l'équation 1 n'implique que les variables Y, X et Z. L'équation 1 peut être généralisée et reformulée en utilisant $X_S = \{X\}$, $X_{\bar{S}} = Z$:

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

Pour calculer cette quantité à partir de données d'observation, il convient de construire un modèle probabiliste f de Y en sachant seulement $\mathbf{X} = X_S \cup X_{barS}$, puis de s'appuyer sur une intégration de Monte-Carlo sur les données d'apprentissage où la probabilité $P(Y|\mathbf{X})$ est estimée par $f(\mathbf{X})$:

$$P(Y|do(x_S)) \simeq \frac{1}{N}\sum_{i=1}^{N} P(Y|X_S = x_S, X^i_{barS}) \tag{2}$$

$$\simeq \frac{1}{N}\sum_{i=1}^{N} f(x_S, X^i_{barS}). \tag{3}$$

Zhao and Hastie (2019) ont démontré le lien entre la formule du Backdoor et le *Partial Dependance Plot* (PDP).

Étant donné un modèle prédictif $f(\mathbf{X})$, un PDP permet de visualiser et d'analyser la dépendance des prédictions par rapport à une variable d'entrée d'intérêt S (laissons $\bar{S}$ être son complément). Le PDP peut être calculé comme indiqué dans l'équation 4.

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

En effet, l'intégration de Monte-Carlo de l'équation 4 sur les données d'apprentissage a exactement la même équation que l'équation 3.

Ce développement démontre que les connaissances causales préalables orientent vers des sélections de variables pertinentes. Qui permettent l'élaboration de modèles prédictifs, où des outils tels que le PDP acquièrent une signification causale.

Illustration : effet de *Discount* sur *Renewal*

Par construction, le modèle causal générateur de données donne accès au véritable effet causal que pyAgrum peut calculer directement par le biais du do-calculus. Le calcul implique un ajustement du Backdoor avec $\{Usage\}$ comme l'ensemble minimal qui satisfait le critère du Backdoor (voir l'équation 5). En effet, deux ensembles satisfont au critère : $\{Usage\}$ et $\{Usage, Customer\ Profil\}$. Pour le premier ensemble, l'équation 1 devient :

$$P(R|do(D = d)) = \sum_U P(R|D = d, U)P(U) \tag{5}$$

Nous appelons cette valeur l'ACE exact.

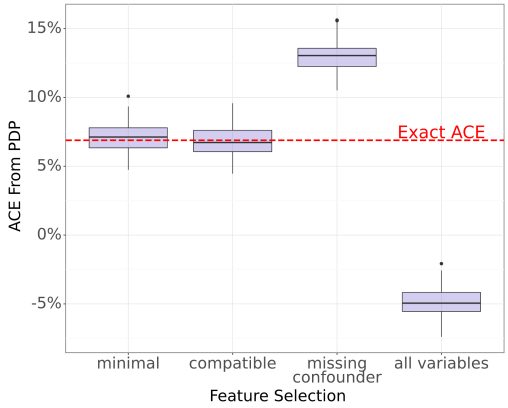

Figure 6: Effet moyen d'une intervention utilisant la méthode PDP pour différentes sélections de variables. L'ACE exact est calculé à l'aide du do-calculus.

Comme indiqué précédemment, l'ajustement du Backdoor peut être estimé à partir d'un échantillon de population à l'aide d'un modèle prédictif formé avec un algorithme standard tel que XGBoost. Le calcul implique une intégration de Monte-Carlo sur un échantillon de population de taille N.

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

$$\simeq \frac{1}{N}\sum_{i=1}^{N} f(D = d, U)$$

f est un modèle de classification formé pour estimer la probabilité de *Renewal* conditionnellement à *Discount* et *Usage*. f est appliqué à un échantillon de population, en prenant *Usage* dans les données et en forçant *Discount* à la valeur d, conformément à la technique PDP.

Nous comparons ensuite l'ACE exact avec les estimations de 100 échantillons de population de taille $N = 10000$. Pour chaque échantillon de population, nous avons formé quatre modèles prédictifs impliquant différentes sélections de variables :

- *minimal* : un ensemble minimal de variables satisfaisant au critère du Backdoor, ici {*Discount*,*Usage*},
- *compatible* : un ensemble plus large de variables compatibles avec le critère du Backdoor, en ajoutant {*Customer Profil*} à l'ensemble minimal,
- *missing confounder* : un ensemble de variables qui ne satisfait pas au critère du Backdoor parce qu'il exclut une variable nécessaire pour bloquer un chemin entre l'action et la cible, en excluant ici *Usage* de l'ensemble *compatible*,
- *all variables* : l'ensemble de toutes les variables connues, incompatible avec le Backdoor Criterion car il contient une conséquence de l'action, à savoir *Visits*.

La technique PDP est ensuite appliquée pour estimer l'effet moyen sur les prédictions d'une intervention entre *Discount=0* et *Discount=1*.

La figure 6 présente les résultats expérimentaux. Les sélections de variables *minimal* et *compatible* fournissent toutes deux une estimation précise de l'effet causal moyen pour *Discount*. En revanche, les deux sélections de variables incompatibles avec le critère du Backdoor conduisent à des estimations sensiblement différentes. Le calcul effectué à partir du modèle avec un facteur de confusion manquant surestime l'effet causal. Il convient de mentionner qu'ici, avec l'utilisation classique de l'ensemble des variables connues, l'estimation et l'ACE sont opposées.

Quantifier une intervention dans un contexte

Une autre question causale pertinente consiste à estimer l'effet d'une intervention dans un contexte spécifique. Pour une intervention sur une variable binaire *X* dans un contexte défini par l'ensemble des variables *Z*, le problème est d'estimer un *uplift* à partir de données d'observation (Rubin 1974; Gutierrez and Gérardy 2017) :

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

L'uplift exacte à l'aide du do-calcul

Selon la règle 2 (échange action/observation) du do-calcul (Pearl 2012) :

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

où $G_{\overline{T_1}\underline{T_2}}$ est le graphe causal obtenu en supprimant toutes les flèches pointant vers les nœuds de T_1 et toutes les flèches émergeant des nœuds de T_2. En remplaçant (T_1, T_2, K) par $(\emptyset, X, Z)$:

Propriété *(Estimation de l'effet de l'intervention) :*

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

où $G_{\underline{X}}$ est le graphe causal obtenu en supprimant toutes les flèches émergeant de X.

En particulier, si Z satisfait au critère du Backdoor par rapport à la paire (X, Y), alors les variables de Z bloquent tous les chemins reliant X à Y qui contiennent une flèche vers X, et la suppression des flèches émergeant de X garantit que $(Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}}}$. Ainsi, si l'ensemble des variables Z satisfait au critère du Backdoor par rapport à la paire (X, Y), nous pouvons estimer l'effet d'une intervention sur X en utilisant directement les probabilités conditionnelles estimées à partir des données d'observation : $P(Y|do(X = x), Z) = P(Y|X = x, Z)$.

Appliquée à notre exemple (figure 3), cette propriété indique que l'uplift de *Discount* peut être correctement estimée si X=*Discount*, Y=*Renewal* et Z vérifie la propriété 8. Dans le cadre d'une analyse d'uplift, Z peut être maximisé, c'est-à-dire que Z ={*Économie, utilisation, profil du client*}.

Estimer à partir d'un échantillon

Dans la modélisation de l'uplift, l'ensemble des variables comprend le traitement (X) et son contexte (Z). Plusieurs techniques de modélisation d'uplift peuvent être appliquées pour estimer une remontée à partir d'un ensemble de données d'observation. Dans l'approche "à deux modèles", des modèles distincts sont ajustés sur les sous-populations de contrôle ($X = 0$) et traitées ($X = 1$), comme dans l'équation 9. Dans l'approche "modèle unique", un estimateur est formé sur l'ensemble de la population, le traitement attribué faisant partie de l'espace variable, comme dans l'équation 10.

$$P(Y|do(X=x),Z) \simeq f_x(Z) \quad (9)$$
$$P(Y|do(X=x),Z) \simeq f(X=x,Z) \quad (10)$$

La sous-section a démontré que les estimations des équations 9 et 10 sont pertinentes si la propriété 8 s'applique.

Illustration : uplift de *Discount*

Puisque nous connaissons le modèle causal de notre processus de génération de données synthétiques, nous pouvons calculer l'uplift exacte à partir de l'équation 6 à l'aide de pyAgrum. D'autre part, la propriété 8 fournit une estimation à partir d'un modèle prédictif formé sur un échantillon de population :

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

La figure 7 compare l'uplift exacte calculé sur le modèle causal de génération de données, avec les estimations de 100 modèles de classification formés sur des populations d'échantillons de 50000 observations. Le graphique de gauche représente les uplifts pour les *clients professionnels*, et le graphique de droite concerne les *clients particuliers*. Les points bleus représentent l'uplift exacte. Les boxplots représentent les uplifts prédits, en vert pour la bonne sélection de variables (D, U, C, E) et en rouge pour une

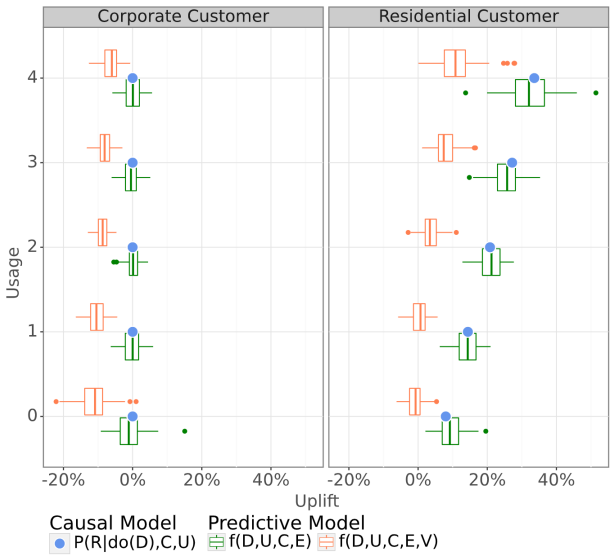

Figure 7: Uplift théorique et uplift du modèle prédictif d'une intervention sur *Discount*.

sélection comprenant toutes les variables connues. Nous observons que pour les *clients professionnels*, les uplifts estimés à l'aide de la sélection correcte sont proches de 0, quelle que soit *Usage*. Cela correspond à la vérité de terrain, où l'uplift est précisément nulle. Les uplifts estimées pour les *clients privés* s'alignent également sur le modèle de génération de données causales. Cependant, avec les approches prédictives classiques utilisant l'ensemble des variables connues, les estimations de l'uplift sont loin des valeurs exactes et peuvent même être opposées. Une fois de plus, les effets causaux estimés à partir d'un modèle prédictif sont assez précis tant que les variables ont été soigneusement (et causalement) sélectionnées.

Conclusion

La principale contribution de cet article est une nouvelle approche XAI qui permet de mieux quantifier les effets de causalité à partir de données d'observation. Nous montrons que l'application directe des outils XAI à un modèle formé à partir de toutes les variables connues sans tenir compte de la causalité peut conduire à des interprétations erronées. Pour résoudre ces problèmes, nous proposons un nouveau cadre pour analyser chaque requête causale séparément sur la base de la structure causale. Cela permet d'obtenir un modèle et une technique d'interprétation sur mesure, fournissant des estimations numériques des effets de causaux. En contrepartie, la réponse à de multiples requêtes causales peut nécessiter l'apprentissage de plusieurs modèles prédictifs.

Dans la communauté XAI, un débat existe autour des notions *true to the model* et *true to the data* (Chen et al. 2020). De ce point de vue, il nous semble que l'assouplissement de la contrainte d'un modèle prédictif préexistant permet à la XAI d'être plus fidèle *aux données*.

Dans notre approche, la causalité guide à la fois la construction et l'analyse des modèles prédictifs. Cependant, il est difficile (voir impossible) de trouver la structure causale complète. Différentes méthodes peuvent être utilisées pour trouver un graphe causal partiellement dirigé (PDAG). Elles peuvent être divisées en deux familles : les méthodes basées sur l'indépendance conditionnelle (Spirtes et al. 2002; Louis et al. 2017; Glymour, Zhang, and Spirtes 2019), et les méthodes basées sur le score (Chickering 2002). Par conséquent, la prochaine étape pour rendre notre approche plus opérationnelle consisterait à étudier comment une connaissance partielle du graphe causal peut être suffisante pour guider la modélisation prédictive et répondre avec précision aux requêtes causales.

Remerciements

Ce travail a été effectué dans le cadre d'une thèse CIFRE (no2020/1640) soutenue par SAP France et l'ANRT (Association Nationale de la Recherche et de la Technologie).

References

Breiman, L. 2001. Random Forests. *Machine Learning* 45(1): 5–32. ISSN 1573-0565.

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, H.; Janizek, J. D.; Lundberg, S.; and Lee, S.-I. 2020. True to the Model or True to the Data?

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.

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. doi:10.1089/big.2017.0104.

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. Skørping, Denmark. URL https://hal.archives-ouvertes.fr/hal-03135721.

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. ISSN 1664-8021.

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. ISSN 10618600, 15372715.

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. doi:10.48550/ARXIV.2302.00736.

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. ISSN 1664-8021.

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. ISBN 9780974903989.

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. doi:10.1037/h0037350.

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*, 307–317. Princeton University Press.

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. ISSN 0735-0015.