Causality-Guided Interpretation of Predictive Models

Author names withheld

Editor: Under Review for CLeaR 2023

Abstract

Recent Machine Learning (ML) methods are increasingly sophisticated and generally improve the accuracy of the models constructed, but at the expense of greater difficulty of interpretation. Moreover, the interpretability of these models is an increasingly sensitive issue in many fields. Indeed, the use of these models in the context of automatic decision-making requires a detailed knowledge of behaviour in order to be able to justify the decision (for example, in the medical domain of automatic prescription, in the legal domain or in a legal context). Practitioners are often interested in causal insights into the underlying data-generating mechanisms, which Machine Learning methods do not generally provide. It is argued here that the search for an answer to a question that involves causality must first involve reflection on the question. The question and the causal framework for answering it must be clearly stated. And otherwise, the abrupt use of widespread explainability tools without any consideration for the causal aspect of the problem can be misleading. Our proposal in this paper is that upstream causal knowledge makes these tools much more relevant.

Keywords: Causality, Machine Learning, Interpretability, XAI

Introduction

Recent Machine Learning (ML) methods are increasingly sophisticated and generally improve the accuracy of the models constructed, but at the expense of greater difficulty of interpretation. Moreover, the interpretability of these models is an increasingly sensitive issue in many fields (Burkart and Huber, 2021). Indeed, the use of these models in the context of automatic decision-making requires a detailed knowledge of behavior in order to be able to justify the decision ; for instance, in the medical domain of automatic prescription, in the legal domain or in a legal context (Rieg et al., 2020). Practitioners are often interested in causal insights into the underlying data-generating mechanisms, which Machine Learning methods do not generally provide. Common causal questions include the identification of causes and effects, predicting the effects of interventions, and answering counterfactual questions.

This article addresses the question of quantifying causal effects from observational data, using off-the-shelf supervised learning algorithms and Interpretable ML techniques. In practice, errors may come from the learned models, which may be biased or suffer from some alteration. To set aside these issues, we give ourselves a **causal model** to define the ground truth. We will thus be able to formally define a Bayes optimal classifier, as a probabilistic model that makes the most probable prediction for a new instance (Mitchell, 1997). By generating large amounts of training data from few unique combinations of input features, the trained experimental classifier models are close to the Bayes optimal classifier. The other benefit of using a causal model as the ground truth is that we will be able to quantify the exact causal effect of some input feature on a target variable using analytical methods such as do-calculus (Pearl, 2000). Thus, we will be in position to analyze explanations about an experimental classification model and verify their fidelity to the generating model.

Considering a classification task, the **machine-learning model** is defined as a real-valued function f that takes a vector of real-valued features as input and the function models the score of a class. Let be a prediction problem of class C, learned from a database composed of N features $\mathbf{X} = \{X_1, X_2, \cdots, X_j, \cdots, X_N\}$ and D rows, the function computing the predicted value according to these variables will be noted $f(X_1 \cdots X_N)$.

Provided with this model, our interest is the **Causal Effect** of an actionable feature on the target. An actionable variable is a variable that can be acted upon in the "real world", i.e. a variable on which one can act and thus control the value.

The **Causal Effect** of A on Y is given by $P(Y|do(A = a))$ (Pearl, 2012) which means the distribution of Y given that we force A to be a (intervention on A). This defines the basis of the examination of causal effects. But quantifying the causal influence of A on Y is a non-trivial question. In order to measure the causal strength, we rely on the Average Causal Effect (ACE) (Janzing et al., 2013) for binary A:

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

This is similar to the concept of average uplift, which is appreciated for its simplicity of understanding and thus facilitates decision-making. Within this framework, a desired property would be to have a non-zero contribution to a variable only if the variable has a causal effect. Another would be that the contribution follows the trend of the prediction. In other words, if the probability of obtaining Y increases when acting on A, the contribution to the variable is positive and in the opposite case, if acting on A decreases the probability of Y, we assign a negative contribution to our actionable variable.

Several tools have been developed to attribute to input features the prediction made by a model. For example, the *Partial Dependence Plots* (PDP) propose to examine the effect of the j-th variable by studying the average prediction when this j-th variable is perturbed (Friedman, 2001) by mixing the values of the j-column in the base, for instance.

The *Individual Conditional Expectation Plots* (ICE) are based on the same idea as the PDPs; but correspond to the study of the prediction f from a given example when the j-th variable is modified (Goldstein et al., 2015) . Thus, the average of all the ICEs corresponds to the PDP.

Another idea for explainability is to provide an importance score for each variable. Breiman (2001) proposes to exchange a variable against noise and then to look at how the prediction f is impacted. Other methods still exist, but each one brings a certain number of constraints that must be respected in order to ensure the relevance of the interpretation (Hooker and Mentch, 2019).

This article will often refer to Shapley values (Shapley, 1953). Several variants (Sundararajan and Najmi, 2020; Frye et al., 2020; Heskes et al., 2020; Wang et al., 2021) have been proposed to address different concerns, but we will focus on the widespread SHAP values as a de facto standard for off-the-shelf Interpretable ML (Lundberg and Lee, 2017).

Shapley values are a method to spread credit among players in a "coalitional game" (von Neumann and Morgenstern, 1947). Here v is a value function associating a real number $v(S)$ to any coalition $S \subseteq \mathbf{X}$. For its use in the field of explicability, a parallel is drawn between prediction and the value function in a game. One can see the input feature X as the players that collaborate to gain f(X), so Shapley values become a good way to explain a model and spread in the area of machine learning (Strumbelj and Kononenko, 2010; Lundberg and Lee, 2017)).

The explanation provided by SHAP values is an excellent basis for understanding the behavior of a predictive model. SHAP values offer an explanation that fits various types of models and are

based on solid mathematical foundations. However, the problem of explainability often lies more in prescribing than in predicting. Predictive Analytics aims at answering questions such as "What is the likely value of Y if I observed X?" and "What are the weights of the evidence leading to this prediction?". On the other hand, Prescriptive Analytics aims at answering questions such as "What intervention should I do to improve Y?" and "When and why should I make such an intervention?". SHAP values quantify the contributions from evidence to the outcome of a predictive model, and thus fit the needs of Predictive Analytics. However, the contributions from evidence can be easily confused with the contributions from interventions. The latter are needed for Prescriptive Analytics. In the following, we present an illustrative example that gave us the intuition of the importance of the causal framework. Then some experiments in which causal analysis would be necessary for the proper application of the interpretability tools: first in the feature selection, then in the estimation of an average causal effect and finally in the estimation of an uplift.

1. Sensitivity to Feature Selection

To illustrate our point, we will use a synthetic example designed using pyAgrum, a library for graphical models (Ducamp et al., 2020). To simplify the reasoning, we create a database with a semantic meaning.

Let's take the task of predicting if the customer is going to renew his cell phone subscription. In order to achieve our assignment, we have access to different features:

- the profile of the client (i.e. is the client's contract a professional subscription) represented by the variable *corporate customer* (aka *profile*),
- the usage of the client as *yearly consumption* (aka *consumption*)
- an offer granted to the customer illustrated by *coupon*,
- the *loyalty* of the client, which typically cannot be directly observed and will be handled as a latent factor,
- *recent visit* that indicates if the customer has visited the provider website recently,
- finally *renewal* that informs on subscription renewal and will be the target for binary classification models

In order to limit the size of the feature space and to train classifier models close to the optimal Bayes model, most of the variables are binary except *yearly consumption* which can take five distinct values. The Figure 1 translate the situation into a graph.

To study this case, we train predictive models using a well-known ML algorithm; XGBoost (Chen and Guestrin, 2016). We train two models on two different sets of features (but same data points); a first model is trained on all known features (i.e. all except the target and the unobserved *loyalty*) and a second is trained after withdrawing an additional variable: *recent visit*. For these two models, we use the SHAP library ((Lundberg and Lee, 2018)) to extract explanations and try to draw conclusions on how to improve customer retention.

The results of the study are given in Figure 2 and Figure 3. The two plots represents, as explained in the documentation of the SHAP library, the SHAP values of every feature for every sample. The plot sorts features by the sum of SHAP value magnitudes over all samples, and uses SHAP values

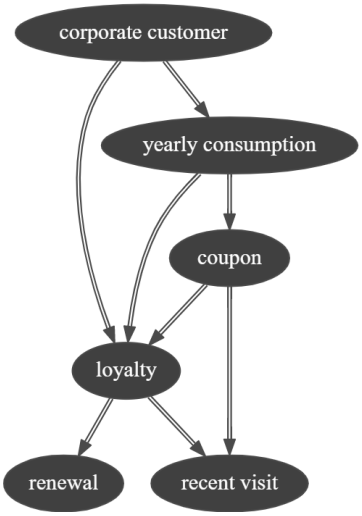

Figure 1: Causal Model used from generating the dataset and thus representing the underlying causality of the data.

to show the distribution of the impacts each feature has on the model output. The colour represents the feature value (red high, blue low).

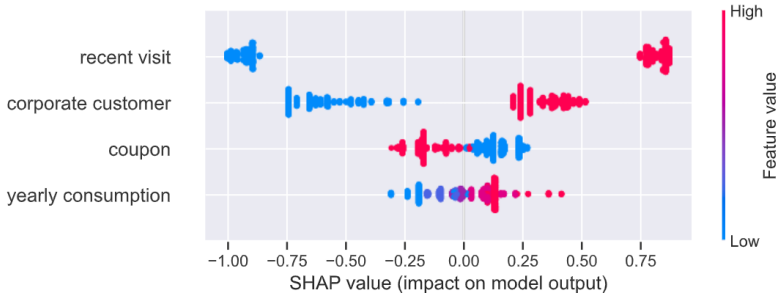

Figure 2: Summary Plot from SHAP, explaining a model trained on all variables

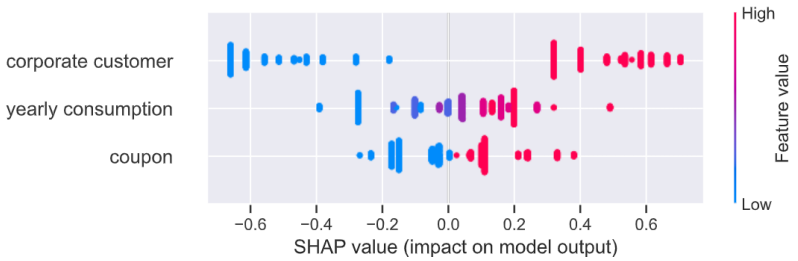

Figure 3: Summary Plot from SHAP, explaining a model trained excluding *recent visit*

A reading of the first graph, Figure 2, suggests that granting a *coupon* would have a negative effect on the target. That is, giving a discount to a customer decreases his chances of renewal. This

seems rather counter-intuitive, as one would tend to think that the granting of a promotion would build customer loyalty and encourage them to renew their subscription. This logic can be seen in the graph in Figure 3, where coupons now appear to have a positive effect on subscription renewals. Given these two opposing explanations, it is clear that the widespread SHAP interpretation method is sensitive to variable selection, and may provide conflicting insights when applied to different models trained from the same data set.

A question naturally arises: can we train and interpret a supervised learning model to get the correct answer to a specific causal question?

2. Estimating An Average Total Causal Effect

The first question we address is the estimation of the average total causal effect of a variable, previously introduced in Equation 1.

2.1. Solution Using Do-Calculus

Within a causal model, the do-calculus framework gives us multiple techniques such as front-door or back-door adjustments to estimate this causal effect. Back to our example (Figure 1), the backdoor adjustment (Pearl, 2000) is suitable for quantifying the causal effect of *coupon* on *renewal*. This adjustment involves a Criterion 2.1 to define a set of variables that should be considered.

The 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.

And 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 **Backdoor Adjustment**:

$$P(Y|do(X = x)) = \sum_z P(Y|X = x, Z = z)P(Z = z) \tag{2}$$

With respect to this criterion, we obtain a minimal set reduced to a singleton variable $\{$*yearly consumption*$\}$ and another set $\{$*yearly consumption*, *corporate profile*$\}$ that is not minimal but still satisfies the criterion.

2.2. Estimates From A Sample Population

Now in possession of a proper manner to estimate the causal effect through the backdoor adjustment, we can observe that the backdoor adjustment formula can be estimated using the same calculations involved in a classic interpretation method for predictive model. This widely used method is the Partial Dependence Plot (PDP) (Friedman, 2001). Given a predictive model f, it is used to visualize and analyse interaction between the target and a set of input features of interest S. It can be computed as shown in Equation 3.

$$f_S = E_{X_{\bar{S}}}[f(x_S, X_{\bar{S}})] = \int f(x_S, X_{\bar{S}})dP(x_{\bar{S}}) \tag{3}$$

Actually, computing a **PDP** often relies on a Monte-Carlo integration over the training data :

$$\hat{f}_S = \frac{1}{N}\sum_{i=1}^{N} f(x_S, X_{\bar{S}}^i) \tag{4}$$

where $\{X_{\bar{S}}^1, X_{\bar{S}}^2, X_{\bar{S}}^3, ..., X_{\bar{S}}^N\}$ are the values of $X_{\bar{S}}$ in the training data. We can observe that the Equation 3 of PDP matches the backdoor equation of Pearl's do calculus (Pearl, 2012), if the set of input variables is compatible with the backdoor criterion (Equation 5).

$$P(Y|do(X_S = x_S)) = \int P(Y|do(X_S = x_S, X_{\bar{S}} = x_{\bar{S}}))dP(x_{\bar{S}}) \tag{5}$$

A more detailed proof of this relationship can be found in (Zhao and Hastie, 2019). This example demonstrates that placing ourselves in a causal framework allowed us to use causal theory to guide our analysis and use well-known tools initially designed for predictive models. However, the results of explainability methods are not always in line with causality, and for good reason, as they do not take into account the constraints of the causal framework, such as selecting a specific set of variables compatible with the effect to be quantified.

In our example of subscription renewal, the causal model gives us access to the true causal effect. This effect can be calculated directly with the pyAgrum library, which implements the Do-Calculus. The calculation involves a backdoor adjustment with { *consumption* } as the minimal set that satisfies the backdoor criterion:

$$P(renewal|do(coupon = c)) = \sum_{consumption} P(renewal|coupon = c, consumption)P(consumption) \tag{6}$$

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

$$P(renewal|do(coupon = c)) \approxeq \frac{1}{N}\sum_{i=1}^{N} P(renewal|coupon = c, consumption) \tag{7}$$

$$P(renewal|do(coupon = c)) \approxeq \frac{1}{N}\sum_{i=1}^{N} f(coupon = c, consumption) \tag{8}$$

Here f is a classifier model trained to estimate the probability of *renewal* conditional on *coupon* and *consumption*. f is applied to a sample population, taking *consumption* from the population and forcing *coupon* to the value c, as per the Partial Dependence Plot technique.

2.3. Experiment

The exact do-calculus calculation from the data generation model gives an average causal effect close to 7%[1].

1. Figures and models can be found in https://github.com/mahdihadjali/CausalityForInterpretingModel

We compared this figure with estimates from several sample populations of size 10000 (a typical size for simple classification tasks). For each sample population, we trained four predictive models involving different feature selections:

- *minimal*: a minimal set of features that satisfies the backdoor criterion, here { *coupon*, *yearly consumption* },
- *compatible*: a larger set of features that is still compatible with the backdoor criterion, here adding *corporate customer* 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 *yearly consumption* from the *compatible* set,
- *all variables*: the set of all known features, that is not compatible with the backdoor criterion because it contains a consequence of the action, namely *recent visit*.

The classifiers were then interpreted using the Partial Dependence Plot technique, to estimate the average effect on predictions of an intervention from *coupon=0* to *coupon=1*.

The plot, Figure 4, represents the experimental results. We observe that both the *minimal* and *compatible* feature selections provided an accurate estimate of the Average Causal Effect for *coupon*. On the other hand, the two feature selections that were not compatible with the backdoor criterion led to significantly different estimates. In other words, whenever the feature selection was incompatible with the backdoor criterion, an intervention on the model inputs was not representative of an intervention on the real world.

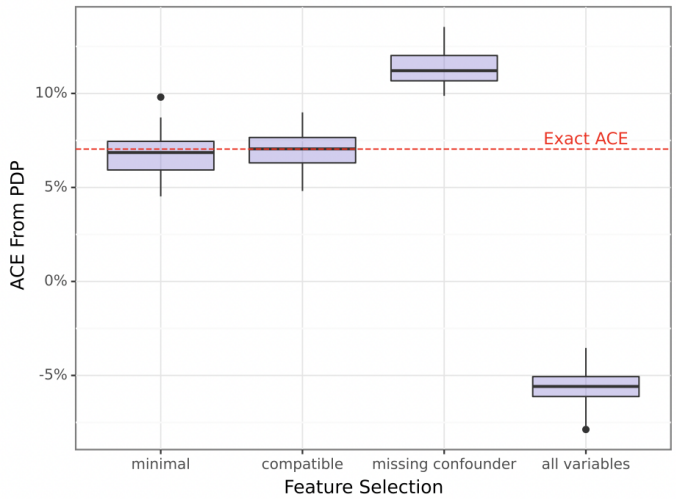

Figure 4: Average Effect of an Intervention, using PDP method

2.4. Relation with the SHAP Framework

Multiple variants of the SHAP framework have been proposed since the original proposal from (Lundberg and Lee, 2017). At least one variant involves a calculation that is similar to the Monte-Carlo integration of the backdoor adjustment: the main effect in the Shapley-Taylor interaction

index (Sundararajan et al., 2020). This technique extend the attribution problem to feature interactions. The notion of Shapley interaction was introduced by (Owen, 1972) but only to analyze pairwise interactions between players. Then it was generalized to study interactions of higher orders (Grabisch and Roubens, 1999). The Shapley-Taylor interaction index method, as its name suggests, is inspired by how Taylor's series decomposes the function value at a certain point using its derivatives at a different point. The Shapley-Taylor interaction index at order 2 is a matrix where the diagonal represent the main contributions of the input features, whereas the other cells represent interaction effects for all pairs of features. The formula for the main effect of feature *i* is then:

$$\phi_{i,i} = F(\{i\}) - F(\emptyset) \tag{9}$$

where F is a characteristic function of the model to be interpreted. If we choose to define F as an Interventional Conditional Expectation as defined in (Lundberg et al., 2020), then $F(\{i\})$ is equivalent to the Monte-Carlo integration over a sample population of Equation 3, and we get, when the feature selection satisfies the backdoor criterion for variable *i*:

$$\phi_{i,i} \approxeq P(Y|do(x_i)) - P(Y) \tag{10}$$

The Shapley-Taylor interaction index at order 2 estimates the interactions between input features. As in the open-source shap library, the contributions are estimated as log-odds rather than probabilities, but we nevertheless verified that if a classifier is trained on a feature selection compatible with the backdoor criterion, then the main contribution calculated for the variable *coupon* does not depend on other variables and is consistent with its expected positive average causal effect.

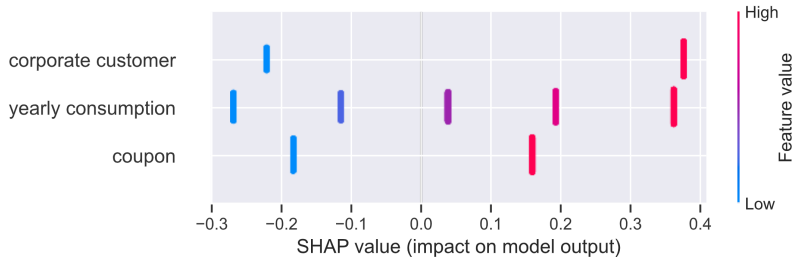

Figure 5: Main contributions from Shapley-Taylor Interaction Indexes (made with SAP Automated Predictive Library)

Then the link between predictive tools and causality does exist. But in order to rely on it, it is necessary to take causality as a whole and therefore its specific requirements. Indeed, if one uses interpretability tools without any consideration for causality, a risk of inconsistency or irrelevance emerges : the results given by the tools will not answer the causal question being considered.

3. Estimating An Uplift

Another relevant causal question is to estimate the effect of an intervention in a specific context. For an intervention on a binary variable *X* knowing a setting defined by the variables *Z*, the problem is to estimate an uplift from observational data:

$$uplift = P(Y|do(X=1), Z) - P(Y|do(X=0), Z) \tag{11}$$

3.1. Naive Approach Using SHAP Values

Let us return to our customer retention example (Figure 1) to formulate the new question. Now that we know the granting of a *coupon* has a positive average effect, we would like to target the customers who are most influenced by this discount. We learned from the previous question that a predictive model should be trained on a selection of features that are compatible with the backdoor criterion, so we start by training such a model, then we extract SHAP values for a test population. If we segment customers according to their profile (corporate vs non-corporate customers), the SHAP analysis in Figure 6 and Figure 7 indicates that granting a *coupon* always has a positive contribution to the predicted target. This contradicts the causal model of our data generator, which was specifically designed so that corporate customers are not influenced by coupons. The intuition from SHAP values is also at odds with what-if simulations using the predictive model to test the impact of an intervention on *coupon*.

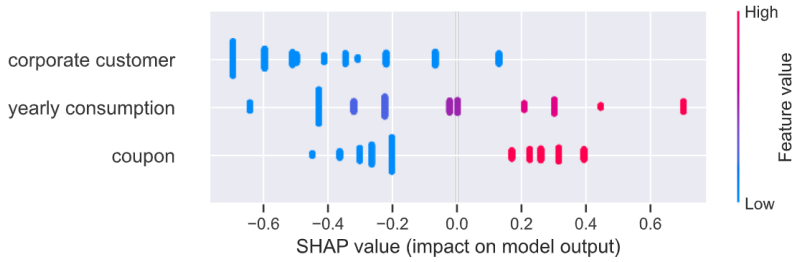

Figure 6: SHAP Summary plot for non-corporate customers

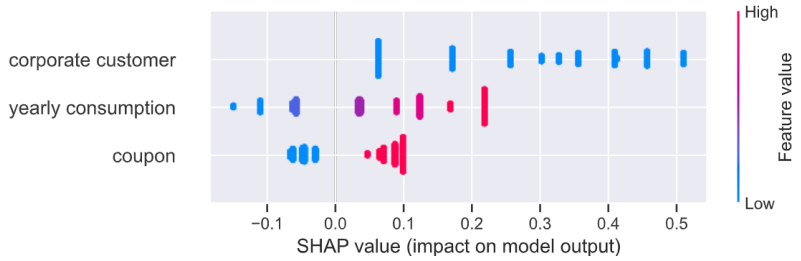

Figure 7: SHAP Summary plot for corporate customers

3.2. Exact Calculation Using The Causal Model

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

$$P(Y|do(X), do(Z), W) = P(Y|do(X), Z, W) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{12}$$

where $G_{\overline{X}\underline{Z}}$ is the causal graph obtained by removing all arrows pointing to nodes in X and all arrows emerging from nodes in Z. If we substitute X with $\emptyset$, Z with X and W with Z, rule 2

becomes:

$$P(Y|do(X), Z) = P(Y|X, Z) \text{ if } (Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}}} \tag{13}$$

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

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

$$P(Y|do(X = x), Z) = P(Y|X = x, Z) \tag{14}$$

In other words, naive what-if simulations of interventions on X using a predictive model of Y provide valid estimations of the effect of interventions in the real world, provided that the predictive model has been trained on a feature selection that satisfies the backdoor criterion for the actionable variable and the target.

An application to the customer retention example gives:

$$P(renewal|do(coupon), consumption, profile) = P(renewal|coupon, consumption, profile) \tag{15}$$

The content of the table corresponds to the uplift equation 15. The formula is estimated by pyAgrum on the causal data generation model (pyAgrum automatically calculates this formula), the results are given in Table 1.

The difference between private and corporate customers can be clearly seen here; coupon has a positive effect on private customers and no effect on corporate customers. Causal analysis therefore points us to the right reasoning, whereas SHAP results, without taking causality into account, lead to the wrong conclusion.Indeed, the *coupon*'s SHAP coalitions involve a mix of corporate and non-corporate clients, and this creates a difference between *coupon=0* and *coupon=1*, even for an observation like *corporate customer=1*.

corporate customer	yearly consumption	uplift
0	0	0.000
	1	0.000
	2	0.000
	3	0.000
	4	0.000
1	0	0.008
	1	0.144
	2	0.208
	3	0.272
	4	0.336

Table 1: Theoretical uplift from an intervention on Coupon

3.3. Relation with Uplift Modelling

Let's consider an additional example: the tax payment problem. A classifier model can predict which individuals are most likely to pay after a reminder, and then this population can be targeted by a mail campaign. Unfortunately, this is not necessarily the optimal decision. Some would have paid regardless of the campaign, targeting them resulted in unnecessary costs. Others were actually going to pay on time from fear of a huge penalty, but the action makes them realize that the actual penalty is much smaller than expected. This problem is often raised in real case studies, for instance in the marketing field (Radcliffe and Surry, 2012; Hansotia and Rukstales, 2002); sub-optimal targeting can lead to losses or even churn. To address the problem, a well studied approach is to turn to uplift modelling (Radcliffe and Surry, 1999). This technique is a branch of machine learning that tries to forecast class probability differences between a group exposed to some action and a control group.

The main requirement of uplift modelling is to have a control population that received a treatment and a test population that did not receive it. Ideally, both populations come from a Randomized Control Trial (RCT), but in practice they may come from so-called "natural experiments". For instance, if for whatever reason all individuals in a US state received an incentive whereas no individual in a neighbouring state received it, and if we assume that the two states have similar populations, then the two populations might be used as test and control sets. The most convenient to conduct this uplift study would be to use observational data and avoid going through the time-consuming and costly RCT process. A benefit of causal analysis could be to detect potential uplift from pure observational data (data where control and test populations were not clearly separate, and where treatment might have both causes and consequences in other variables), by using *do-calculus* theory (Pearl, 2012).

3.4. Experiments

In practice, several techniques from uplift modelling are applicable to estimate the uplift on our observational dataset: the two-model approach (a separate model fitted to the control and treatment groups) or the unified model (a single model with the allocated treatment being part of the feature space). We will use the latter. As in section 2.3, we compare the uplift of the causal model with an uplift given by classification models trained on a sample database of 50000 observations. The predicted uplift, shown in blue, in the Figure 8, represents the distribution of outcomes from 100 XGBoost models. In red, we have the uplift from the ground truth model (i.e. the causal model).

The Figure 8 consists of two parts; the left-hand side represents the uplift for a corporate customer and the right-hand side the uplift for a private customer. We observe that the estimated uplift for a corporate customer is very small, regardless of the yearly consumption, and in line with the ground truth where the uplift is exactly zero. On the right-hand side, the estimated uplifts for private customers are also in line with the causal data generation model.

3.5. Relation with the SHAP Framework

Although it might be possible to design an asymmetric variant of the SHAP framework such that the SHAP value ϕ_i of a variable i is estimated as the marginal effect of an intervention on i in a context set by the other known features, we can observe that SHAP calculations are typically CPU-intensive, whereas we just showed that an uplift can be directly estimated by a straightforward application of a predictive model trained on a proper set of features. We can also observe that the asymmetry comes neither from the data nor from the trained classifiers, but rather from the question: uplift modelling

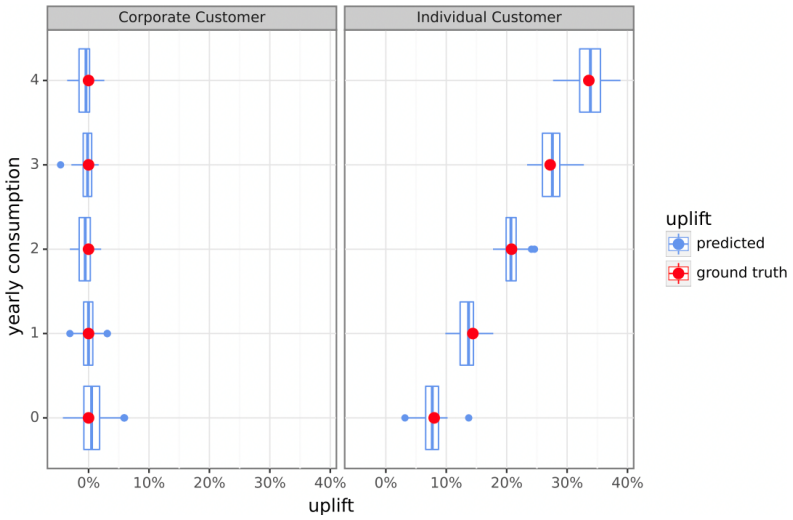

Figure 8: Predicted & Theoretical uplift from an intervention on Coupon

techniques are typically applied on data from Randomized Controlled Trials, where the treatment is not causally connected to any other input feature.

4. Conclusion

The core idea of the article is to show how causality can bring useful insights for practitioners when explaining predictive models. Before using the overpowering tools of eXplainable Artificial Intelligenge (XAI), one must clearly define the question one seeks to answer. We have shown that a blunt approach applying XAI tools on a model trained from all known features, with no consideration of causality, can lead to flawed interpretations. On the other hand, a sufficient knowledge of the causal relationships between variables can guide the selection of features prior to training models, and the tools from causal reasoning can drive the choice of XAI techniques that are suitable for the precise question of interest.

In general, the causal framework helps to guide the analysis of predictive models. One difficulty remains: finding the causal graph that describes the causal links between the variables. Different methods can be used to find such a graph, they can be divided into two families: methods based on conditional independence (Spirtes et al., 2000), (Louis et al., 2017), (Glymour et al., 2019) and score-based methods (Chickering, 2002). The results of these methods are not always usable, but a partial causal graph can often be extracted. For future work, we plan to investigate how a partial knowledge of the causal graph may be sufficient to guide the usage of predictive modelling to correctly answer causal questions.

References

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

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

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322.

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

Gaspard Ducamp, Christophe Gonzales, and Pierre-Henri Wuillemin. 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*, pages 609–612, Skørping, Denmark, September 2020. URL https://hal.archives-ouvertes.fr/hal-03135721.

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

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability. NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. ISSN 10618600, 15372715.

Michel Grabisch and Marc Roubens. "an axiomatic approach to the concept of interaction among players in cooperative games". *International Journal of Game Theory*, 28:547–565, 11 1999.

Behram Hansotia and Brad Rukstales. Incremental value modeling. *Journal of Interactive Marketing*, 16(3):35–46, 2002. ISSN 1094-9968.

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

Giles Hooker and Lucas Mentch. Please Stop Permuting Features: An Explanation and Alternatives. *ArXiv*, May 2019. arXiv: 1905.03151.

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324 – 2358, 2013.

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*, 13, 2017. ISSN 1664-8021.

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

Scott M Lundberg and Su-In Lee. Shap. https://github.com/slundberg/shap, 2018.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5):P64–P79, 1972. ISSN 00251909, 15265501.

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

Judea Pearl. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 3–11, Arlington, Virginia, USA, 2012. AUAI Press. ISBN 9780974903989.

Nicholas Radcliffe and Patrick Surry. Differential response analysis: Modeling true responses by isolating the effect of a single action. *Credit Scoring and Credit Control IV*, 1999.

Nicholas J. Radcliffe and Patrick D. Surry. Real-world uplift modelling with significance-based uplift trees. 2012.

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*, 15(12):1–20, 12 2020.

SAP. SAP Automated Predictive Library, 2014. URL https://help.sap.com/docs/apl.

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, Princeton, 1953.

Pater Spirtes, Clark Glymour, Richard Scheines, Stuart Kauffman, Valerio Aimale, and Frank Wimberly. Constructing bayesian network models of gene expression networks from microarray data. 2000.

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

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. 2020.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. 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 9259–9268. PMLR, 13–18 Jul 2020.

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

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.

Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of business economic statistics : a publication of the American Statistical Association*, 2019, 2019. ISSN 0735-0015.