Submitted to *Management Science*
manuscript

The foundations of cost-sensitive causal classification

Wouter Verbeke, Diego Olaya, Jeroen Berrevoets, Sebastián Maldonado
Faculty of Economics and Business, Katholieke Universiteit Leuven, Leuven, Belgium
Faculty of Social Sciences & Solvay Business School, Vrije Universiteit Brussel, Brussels, Belgium
School of Economics and Business, University of Chile, Santiago, Chile
wouter.verbeke@kuleuven.be, diego.olaya@vub.be, jeroen.berrevoets@vub.be, sebastianm@fen.uchile.cl.

Classification is a well-studied machine learning task which concerns the assignment of instances to a set of outcomes. Classification models support the optimization of managerial decision-making across a variety of operational business processes. For instance, customer churn prediction models are adopted to increase the efficiency of retention campaigns by optimizing the selection of customers that are to be targeted. Cost-sensitive and causal classification methods have independently been proposed to improve the performance of classification models. The former considers the benefits and costs of correct and incorrect classifications, such as the benefit of a retained customer, whereas the latter estimates the causal effect of an action, such as a retention campaign, on the outcome of interest. This study integrates cost-sensitive and causal classification by elaborating a unifying evaluation framework. The framework encompasses a range of existing and novel performance measures for evaluating both causal and conventional classification models in a cost-sensitive as well as a cost-insensitive manner. We proof that conventional classification is a specific case of causal classification in terms of a range of performance measures when the number of actions is equal to one. The framework is shown to instantiate to application-specific cost-sensitive performance measures that have been recently proposed for evaluating customer retention and response uplift models, and allows to maximize profitability when adopting a causal classification model for optimizing decision-making. The proposed framework paves the way toward the development of cost-sensitive causal learning methods and opens a range of opportunities for improving data-driven business decision-making.

Key words: Cost-sensitive learning, causal classification, performance, uplift modeling, business analytics

1. Introduction

A range of business decision-making processes across management domains are supported by classification models that are learned from historical data by means of machine learning methods. Classification is a well-studied machine learning task which concerns the assignment of instances to a predefined set of outcome classes. For instance, a customer churn prediction model classifies customers as future churners or non-churners, which allows to select customers that are likely to churn for targeting in a retention campaign (Ben Rhouma and Zaccour 2018, Simester et al. 2020).

Two improvements to standard classification methods have been proposed for improving the performance of classification models:

1. **Cost-sensitive classification methods take into account asymmetric costs and benefits related to incorrect and correct classifications, respectively** (Elkan 2001). For instance, a future churner which is correctly identified by a churn prediction model may be retained by the retention campaign and yield a benefit. Incorrectly classifying a churner as a non-churner, on the other hand, will involve a cost. This cost, moreover, is larger than the cost of incorrectly classifying a non-churner as a churner as it is usually more costly to lose a customer than to target a non-churner in a retention campaign. Cost-sensitive classification methods that take into account these imbalanced costs and benefits better align with the true business objective, i.e., maximizing profit, as shown in literature (Verbeke et al. 2012, Verbraken et al. 2012).
2. **Causal classification methods predict the net effect of an action on an outcome of interest**(Wager and Athey 2018). For instance, uplift models allow to estimate the change in the probability for a customer to churn when targeted with a retention campaign, which aligns with the business objective of maximizing the reduction in churn instead of maximizing the number of correctly identified churners. It has been shown in literature that causal learning for maximizing the effect of marketing campaigns significantly increases the profitability (Ascarza 2018, Devriendt et al. 2019, Gubela et al. 2019).

Whereas cost-sensitive classification methods merely extend upon standard, i.e., cost-insensitive, classification methods, causal classification methods involve a paradigm shift. Essentially, when causal learning is applied for supporting decision-making, the aim is to learn a simulation model from data. Hence, causal learning can be considered to be a type of prescriptive analytics and to be part of the realm of operations research, since the objective is to directly optimize decisions.

Research gap. Both types of methods have been shown in literature to substantially improve the use of classification models for supporting business decision-making (Höppner et al. 2018, Ascarza 2018). However, to the best of our knowledge, no methods have been proposed in literature for combining both improvements and to develop an approach to cost-sensitive causal classification.

Contributions. Therefore, in this article, we define cost-sensitive causal learning as a specialized machine learning task and develop a framework that integrates cost-sensitive and causal classification. The proposed framework formalizes the scattered body of literature on cost-insensitive performance measures and leads to the definition of novel performance measures, both cost-sensitive and cost-insensitive, for evaluating causal classification models. The framework is shown to instantiate to a range of existing performance measures for evaluating conventional classification models.

As such, we proof that classification is to be regarded as a specific case of causal classification. The framework allows to take into account uncertain operating conditions (Hernández-Orallo et al. 2012) and to achieve a robust evaluation, in a manner similar as the H-measure (Hand 2009) and the related Expected Maximum Profit measure (Verbraken et al. 2012, Garrido et al. 2018).

Finally, the practical use of the framework toward business decision-making is illustrated by presenting three cases. We shown that the framework instantiates to two application-specific profit measures that have been recently proposed in literature for evaluating customer churn (Devriendt et al. 2019) and customer response (Gubela et al. 2019) uplift models.

The proposed framework stands on the shoulders of the cost-sensitive and causal learning paradigms. Their integration opens a range of opportunities for developing cost-sensitive causal learning methods as discussed in the conclusions and future research section.

Outline. The remainder of the article is organized as follows. Section 2 discusses a mathematical framework for evaluating classification models, both in a cost-insensitive and a cost-sensitive manner. In Section 3, we extend upon this framework and formalize the evaluation of classification performance in a relative manner, i.e., when compared to a baseline model of choice. Next, in Section 4, we introduce an equivalent framework for evaluating the performance of causal classification models, which allows us to formalize a range of existing performance measures as well as to introduce a novel measure. Then, in Section 5, the framework is further extended to facilitate cost-sensitive evaluation of causal classification models. Section 6 presents three cases to illustrate the practical use of the framework toward business decision-making. Finally, in Section 7, we present conclusions and outline a range of opportunities for further research.

2. Cost-sensitive classification performance

In this section, we introduce a mathematical framework for cost-sensitive performance evaluation of classification models. The mathematical notation is largely adopted from Hand (2009) and Hernández-Orallo et al. (2012).

Classification concerns the assignment of instances x from an instance space $X \in \mathcal{X} \subseteq \mathbb{R}^n$ of dimension n to one of a number of classes y from an output space $Y = \{0, ..., C\}$. In this paper, we focus on binary classification, i.e., $Y = \{0, 1\}$. By convention, we call class 0 the positive class or outcome and class 1 the negative class or outcome. Instances with $y = 0$ are called positive instances and with $y = 1$ negative instances. A classification model is a function $m : X \to S$ which maps instances x to real numbers, i.e., scores $s = m(\mathbf{x})$ in a domain $S \subset \mathbb{R}$. Note that scores can be either unscaled values or calibrated probability estimates, $\hat{p}(y = 1|x)$. We adopt the convention that higher scores express a stronger belief that the instance is of class 1. As such, scores allow to rank instances from low to high probability of being of class 1.

For classifying instances and producing outputs $y \in \{0,1\}$, a classification model can be converted into a classifier by setting a decision threshold t on the scores. Instances with a score below the threshold, $s < t$, are classified in the positive class and instances with a score above the threshold are classified in the negative class.

Classification methods learn a classification model from a data set $\mathcal{D} = \{(x_i, y_i) \ : i = 1, ...N\}$ with $N = |\mathcal{D}|$ the number of instances in the data set. $\mathcal{D}_k$ is the subset of instances in class $k \in Y$, with $N_k = |\mathcal{D}_k|$ the number of instances in class k in data set $\mathcal{D}$, and $\pi_k = N_k/N$ is the prior probability of instances belonging to class k. Since $\sum_k N_k = N$, we have $\sum_k \pi_k = 1$ which for the binary case yields $\pi_0 + \pi_1 = 1$. We call π_0 the positive class proportion and π_1 the negative class proportion. The score density for class k will be denoted by f_k and the cumulative distribution by F_k. Hence, $F_0(t) = \int_{-\infty}^{t} f_0(s) ds = P(s \leq t | y = 0)$ is the proportion of positive instances with a score s below the threshold t.

The confusion matrix, which we denote with $\mathbf{CF}(t)$ or $\mathbf{CF}$ in short, summarizes the number of correctly and incorrectly classified instances of the positive and negative class for a threshold t:

$$\mathbf{CF} := \begin{array}{c c l} & \text{Prediction} & \\ & \begin{array}{cc} \hat{Y}=0 & \hat{Y}=1 \end{array} & \\ & \left[\begin{array}{cc} \pi_0 F_0(t) & \pi_0(1-F_0(t)) \\ \text{(True positives)} & \text{(False negatives)} \\ \pi_1 F_1(t) & \pi_1(1-F_1(t)) \\ \text{(False positives)} & \text{(True negatives)} \end{array}\right] & \begin{array}{l} Y=0 \\ \\ Y=1 \end{array} \quad \text{Outcome} \end{array} \tag{1}$$

The confusion matrix as defined in Equation 1 expresses the number of correctly and incorrectly classified instances for each class as a proportion of the total number of instances N and as a function of the class proportion π_k and the cumulative class distributions $F_k(s)$. For instance, the lower left element is the proportion of all instances that belong to class 1 that are incorrectly classified in class 0 for threshold t. These are called false positives.

Various conventional classification performance measures can be defined in terms of the confusion matrix, e.g.:

$$\text{Accuracy} := \pi_0 F_0(t) + \pi_1 \big(1 - F_1(t)\big), \tag{2}$$

$$\text{Sensitivity} := F_0(t), \tag{3}$$

$$\text{Specificity} := 1 - F_1(t). \tag{4}$$

Note that, since the confusion matrix is a function of the decision threshold, t also the above measures depend on it.

The Receiver Operating Characteristic (ROC) curve plots for all possible thresholds the sensitivity ($F_0(t)$) on the Y-axis versus the false discovery rate ($F_1(t)$) on the X-axis. The ROC curve visualizes the trade-off between both, as achieving an increase in sensitivity typically comes at the price of an increase in false discovery rate. As a measure for assessing the performance of a classification model, often, the area under the ROC curve (AUC) measure is used. The AUC is defined as (Hand 2009):

$$\mathrm{AUC} := \int F_0(s) f_1(s) \mathrm{d}s. \tag{5}$$

The gains curve, which is also denominated the cumulative accuracy profile (CAP) or Lorenz curve, is an alternative to the ROC curve and plots the sensitivity in function of the predicted positive rate, ϕ, which is defined as the proportion of instances that are classified in the positive class:

$$\phi(t) = \pi_0 F_0(t) + \pi_1 F_1(t). \tag{6}$$

The predicted positive rate is a function of the threshold, i.e., $\phi : \mathcal{S} \rightarrow [0,1]$. For brevity, it will be denoted with ϕ. For setting a classification threshold, either a value of the predicted positive rate or a value of the score can be specified.

The Gini coefficient is equal to the ratio of the area in between the gains curve and the diagonal, and the area under the gains curve of the perfect model and above the diagonal, with the diagonal to be interpreted as the gains curve of a random classification model:

$$\mathrm{Gini} := \frac{2\int F_0(\phi)\mathrm{d}\phi - 1}{1-\pi_0}. \tag{7}$$

A third popular method for visualizing model performance is the lift curve, which plots the lift in function of the predicted positive rate. The lift at a predicted positive rate, $\lambda(\phi)$, is defined as the ratio of the proportion of positive outcomes among the predicted positive instances and the overall proportion of positive instances, i.e., the prior positive class probability, π_0:

$$\lambda(\phi) := \frac{\frac{\pi_0 F_0(\phi)}{\phi}}{\pi_0} = \frac{F_0(t)}{\phi}. \tag{8}$$

In practical applications, the four types of outcomes in the confusion matrix, i.e., true and false positives and true and false negatives, may have very different implications. False positive and false negative classifications typically involve different costs resulting from wrongful decision-making, whereas the benefits of true positive and true negative classifications are typically different as well. The costs and benefits associated with the four segments in the confusion matrix can be specified in a cost-benefit matrix, **CB**:

$$\mathbf{CB} := \begin{array}{c} \text{Prediction} \\ \begin{array}{cc} \hat{Y}=0 & \hat{Y}=1 \end{array} \\ \begin{bmatrix} \begin{matrix} b_{(0,0)} \\ (b_0) \end{matrix} & \begin{matrix} c_{(0,1)} \\ (c_0) \end{matrix} \\ & \\ \begin{matrix} c_{(1,0)} \\ (c_1) \end{matrix} & \begin{matrix} b_{(1,1)} \\ (b_1) \end{matrix} \end{bmatrix} \end{array} \begin{array}{l} Y=0 \\ \\ Y=1 \end{array} \quad \text{Outcome} \tag{9}$$

The benefit of correctly classifying an instance from class k is denoted with $b_{(k,k)}$, or b_k in short. The cost of misclassifying an instance from class k to class l is denoted with $c_{(l,k)}$, with $l \neq k$, or c_k in short. By convention, the value of a benefit b_k in the cost-benefit matrix is positive or equal to zero and the value of a cost c_k is negative or equal to zero.

By summing the products of the elements of the confusion matrix and the corresponding elements in the cost-benefit matrix, we obtain the classification profit per instance, which we denote with P:

$$P := \Sigma_{i,j}(\mathbf{CF} \circ \mathbf{CB}), \tag{10}$$

with $\Sigma_{i,j}(\mathbf{A})$ the sum of all elements of matrix $\mathbf{A}$ and $\circ$ the Hadamard or element-wise product.

As a measure for evaluating the performance of a classification model, Verbeke et al. (2012) proposed the maximum profit measure, MP, which is defined as:

$$MP := \max_{\forall t} \big(P(t; \mathbf{CB})\big) = P(t^*; \mathbf{CB}), \tag{11}$$

with t^* the profit-maximizing threshold. As an extension to the MP measure, Verbraken et al. (2014) proposed the Expected Maximum Profit (EMP) measure, which acknowledges that the cost and benefit parameters may not be exactly known or may vary in time or across instances following some joint probability distribution, $h(\mathbf{CB}) = h(b_0, c_0, b_1, c_1)$. The EMP measure is defined as:

$$EMP := \int_{\mathbf{CB}} P(t^*; \mathbf{CB}) \cdot h(\mathbf{CB}) d\mathbf{CB}. \tag{12}$$

The EMP measure achieves a robust evaluation by allowing to take into account random cost and benefit parameters in evaluating a classification model, reflecting the uncertainty on the exact values of these parameters or, when benefits and costs are random, their distribution.

3. Relative cost-sensitive classification performance

In this section, we elaborate on a matter that is, to the best of our knowledge, overlooked in literature and requires formalization, i.e., the implicit choice of a baseline model underlying the evaluation of model performance in a relative manner. The various possible choices and its impact toward the evaluation of performance is of crucial importance in evaluating causal classification models, as will become apparent in Section 5

The performance of a classification model can be measured in absolute sense, for instance by means of P, MP or EMP as defined in Section 2, with a value of zero meaning that instances are not classified and that no outcomes occur. Usually, however, the performance of a classification model is, implicitly, evaluated in a relative manner, i.e., in comparison to some meaningful baseline model. Three common baseline models in literature are (1) a perfect model, which classifies all instances correctly and which is typically adopted in cost-sensitive learning (Elkan 2001); (2) a dummy model, which classifies all instances in one class and which is typically adopted in profit-driven learning (Stripling et al. 2018, Höppner et al. 2018); (3) a random model, which classifies all instances randomly and which is typically adopted in cost-insensitive learning.

For calculating the profit relative to a perfect model, a hollow cost-benefit matrix is used with $b_0 = b_1 = 0$, yielding a zero benefit for correctly classified instances. Hence, only the cost of errors is taken into account in calculating measures such as expected loss (Hernández-Orallo et al. 2012), total misclassification cost ($TMC = N \cdot P$) or average misclassification cost ($AMC = P$) (Verbeke et al. 2017), which is also called loss.

For calculating the profit relative to a dummy model that classifies all instances in the positive class, a cost-benefit matrix is used with the benefit of a true positive and the cost of a false positive equal to zero, $b_0 = c_0 = 0$. Hence, no benefit or cost is taken into account for instances that are classified in the positive class when comparing performance to a model that classifies all instances in the positive class.

No straightforward alteration to the cost-benefit matrix can be made which allows calculating the profit relative to a random model. To this end, an alternative approach is needed.

As an alternative approach for comparing performance to a baseline model by altering the cost-benefit matrix, which appears to be the common practice in literature as well as in software implementations, we define the relative profit, P_r, for calculating the profit relative to a baseline model of choice straightforward as the difference between the absolute profit that is obtained by by the classification model, P, and absolute profit that is obtained by the baseline model of choice, P_b:

Definition 1.

$$P_r := P - P_b \tag{13}$$

If we specify the baseline model of choice in terms of its confusion matrix, $\mathbf{CF}_b$, then we can elaborate the above equation using Equation 10 for calculating the absolute profit of both the classification model, P, and the baseline model, P_b, which yields the following formula for P_r:

$$\begin{aligned} P_r &= \sum_{i,j} (\mathbf{CF} \circ \mathbf{CB}) - \sum_{i,j} (\mathbf{CF}_b \circ \mathbf{CB}), \\ &= \sum_{i,j} \big((\mathbf{CF} - \mathbf{CF}_b) \circ \mathbf{CB} \big), \\ &= \sum_{i,j} (\mathbf{E} \circ \mathbf{CB}). \end{aligned} \tag{14}$$

Definition 2. The difference between the confusion matrix of a classification model, $\mathbf{CF}$, and the confusion matrix of a baseline model $\mathbf{CF}_b$, is the effect matrix, $\mathbf{E}$.

$$\mathbf{E} := \mathbf{CF} - \mathbf{CF}_b \tag{15}$$

THEOREM 1.

$$\sum_{i,j} \mathbf{E} = 0$$

Proof of Theorem 1 By definition, the sum of the elements of a confusion matrix is equal to one, i.e., $\sum_{i,j} \mathbf{CF} = \sum_{i,j} \mathbf{CF}_b = 1$. Hence, we have:

$$\begin{aligned}
\sum_{i,j} \mathbf{E} &= \sum_{i,j} (\mathbf{CF} - \mathbf{CF}_b) \\
&= \sum_{i,j} \mathbf{CF} - \sum_{i,j} \mathbf{CF}_b \\
&= 1 - 1 \\
&= 0 \quad \square
\end{aligned}$$

The baseline models that are frequently used in literature for relative performance evaluation are the following. The baseline confusion matrix for a perfect model, which classifies all instances correctly:

$$\mathbf{CF}_b^{\text{perf}} = \begin{bmatrix} \pi_0 & 0 \\ 0 & \pi_1 \end{bmatrix}. \tag{16}$$

The baseline confusion matrix for a dummy model which classifies all instances in the positive class:

$$\mathbf{CF}_b^{\text{pos}} = \begin{bmatrix} \pi_0 & 0 \\ \pi_1 & 0 \end{bmatrix}. \tag{17}$$

Similarly, for a dummy model which classifies all instances in the negative class:

$$\mathbf{CF}_b^{\text{neg}} = \begin{bmatrix} 0 & \pi_0 \\ 0 & \pi_1 \end{bmatrix}. \tag{18}$$

The baseline confusion matrix for a random model, which classifies instances in a class with a probability equal to the prior class probability:

$$\mathbf{CF}_b^{\text{rand}} = \begin{bmatrix} \pi_0 \cdot \pi_0 & (1 - \pi_0) \cdot \pi_0 \\ (1 - \pi_1) \cdot \pi_1 & \pi_1 \cdot \pi_1 \end{bmatrix}. \tag{19}$$

The baseline confusion matrix for obtaining the absolute profit is the zero matrix:

$$\mathbf{CF}_b^{\text{abs}} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \tag{20}$$

yielding $P_b = 0$ and $P_r = P$.

Although Equation 13 in our view is conceptually a more appropriate approach than adjusting the cost-benefit matrix, the latter approach has a practical advantage in requiring fewer cost-benefit parameters to be specified. Although a full and accurate specification of the **CB** matrix (or the joint probability distribution $h(\mathbf{CB})$ is obviously desirable for an accurate evaluation, in practice it may be difficult or impossible to obtain this information (Hernández-Orallo et al. 2012). The definition of relative profit and of the effect matrix as proposed in this section provides a formal basis for developing a framework for evaluating the performance of causal classification models in the next section.

4. Causal classification performance

In this section, we first introduce causal classification. Then, we extend upon the framework for evaluating the performance of conventional classification models, as introduced in the previous section, for evaluating the performance of causal classification models. Finally, the extended framework is used to formalize existing and introduce novel performance measures for evaluating causal classification models.

4.1. Causal classification

Causal classification concerns the assignment of instances x from an instance space $X \in \mathcal{X} \subseteq \mathbb{R}^n$ to one of a number of classes y from an output space $Y = \{0, ..., C\}$, depending on the applied action w from an action space $W \in \mathcal{W} \subseteq \mathbb{R}$. Causal classification concerns the estimation of the effect of an action on an outcome at the instance level, i.e., of the individual treatment effect (ITE). This causal inference task is encountered in literature under different names, e.g., uplift modeling (Devriendt et al. 2018), heterogeneous treatment effect estimation (Imai et al. 2013), individualized treatment rule learning (Qian and Murphy 2011) and conditional average treatment effect estimation (Shalit et al. 2017).

We focus on binary causal classification, i.e., $Y = \{0, 1\}$, in function of a binary action, i.e., $W = \{0, 1\}$. By convention, we assume that outcomes of the positive class, $Y = 1$, yield a larger benefit than outcomes of the negative class. The purpose of learning causal classification models is to identify instances for which application of the action ($W = 1$) has a positive effect, i.e., causes the outcome to be of the positive class whereas it would be negative if the action is not applied ($W = 0$). For instance, future churners may be retained by targeting them with a retention campaign. Note that applying an action $W = 1$ may as well have no effect or a negative effect, i.e., cause a positive outcome to become negative. For instance, targeting a no-churner may cause the customer to churn.

From a technical perspective, causal classification differs from conventional classification since the effect that we want to estimate, i.e., the ITE, cannot be observed. In real-world applications, only one action at a time can be applied to a subject. Hence, we cannot observe the difference in outcome, i.e., the effect, when applying different actions. This problem is called the fundamental problem of causal inference (Holland 1986) and has given rise to a specialized class of methods for learning causal classification models.

For learning a causal classification model with a binary action variable, two random samples are required which are assumed to be independent and identically distributed and representative for the population of interest. A first sample concerns subjects that have been exposed to the action $W = 1$. This sample is typically called the treatment group and will be denoted with $\mathcal{D}^T$. A second sample concerns subjects that have not been exposed to the action. This sample is typically called the control group and will be denoted with $\mathcal{D}^C$. Then $\mathcal{D} = \{\mathcal{D}^T, \mathcal{D}^C\} = \{(x_i, w_i, y_i) : i = 1, ..., N\}$ with $N = |\mathcal{D}|$ the number of instances in the data set, $N^T = |\mathcal{D}^T|$ and $N^C = |\mathcal{D}^C|$ the number of instances in the treatment and control group, respectively. The number of instances of class k in the treatment and control group is denoted with N_k^T and N_k^C, respectively. The prior class probabilities are denoted with $\pi_k^T = N_k^T / N^T$ for the treatment group, and with $\pi_k^C = N_k^C / N^C$ for the control group; by definition, it holds that $\pi_1^T + \pi_0^T = 1$ and $\pi_1^C + \pi_0^C = 1$. The score density functions of positive and negative class instances in the treatment group are denoted with $f_0^T(s)$ and $f_1^T(s)$, respectively, whereas the cumulative distributions are denoted with $F_0^T(s)$ and $F_1^T(s)$ for the positive and negative class, respectively. For the control group, we have $f_0^C(s)$, $f_1^C(s)$, $F_0^C(s)$ and $F_1^C(s)$.

A binary causal classification model is a function $\dot{m} : X \to S$ which maps instances x to real numbers, i.e., scores $s = \dot{m}(\mathbf{x})$ in a domain $S \subset \mathbb{R}$. Note that scores can be either unscaled values or calibrated estimates of the individual treatment effect. We adopt the convention that a lower score indicates a larger probability for the outcome of an instance to change from the negative to the positive class. As such, scores allow to rank instances from a large positive effect to a large negative effect. The challenge in evaluating causal classification models is to assess the accuracy of this ranking with respect to the effect of the action on the outcome.

To causally classify instances and to decide on the action that is to be applied for individual subjects depending on their score, a decision threshold t on the scores is to be set. A causal classification model with a threshold is called a causal classifier. For subjects with a score below the threshold, $s < t$, the action $W = 1$ is prescribed, i.e., they are classified in the positive class conditional on application of the action. We call this set of instances the action segment. The baseline action $W = 0$, which is typically no action, is prescribed for subjects with a score above the threshold. We call this set of instances the no-action segment. The proportion of positive and negative outcomes in the action and no-action segment is reported in the prescription-outcome matrix.

Definition 3. The prescription-outcome matrix, **PO**, reports the proportion of instances with a positive and a negative outcome in the action and the no-action segment of a sample, as segmented by a causal classification model with threshold t.

We define the prescription-outcome matrix for the treatment group, which we denote with $\mathbf{PO}^T$, as follows:

$$\mathbf{PO}^T := \begin{array}{cc} \begin{array}{cc} \multicolumn{2}{c}{\text{Prescribed action}} \\ W=1 & W=0 \end{array} & \\ \begin{bmatrix} \pi_0^T F_0^T(t) & \pi_0^T\left(1-F_0^T(t)\right) \\ \pi_1^T F_1^T(t) & \pi_1^T\left(1-F_1^T(t)\right) \end{bmatrix} & \begin{array}{l} \text{Y=0} \\ \text{Y=1} \end{array} \ \text{Outcome} \end{array} \tag{21}$$

Likewise, for the control group:

$$\mathbf{PO}^C := \begin{array}{cc} \begin{array}{cc} \multicolumn{2}{c}{\text{Prescribed action}} \\ W=1 & W=0 \end{array} & \\ \begin{bmatrix} \pi_0^T F_0^C(t) & \pi_0^C\left(1-F_0^C(t)\right) \\ \pi_1^C F_1^C(t) & \pi_1^C\left(1-F_1^C(t)\right) \end{bmatrix} & \begin{array}{l} \text{Y=0} \\ \text{Y=1} \end{array} \ \text{Outcome} \end{array} \tag{22}$$

The prescription-outcome matrix strongly resembles the confusion matrix, as defined in Section 2, but contrasts prescriptions versus outcomes instead of predictions versus outcomes.

We proof that the sum of the elements of the prescription-outcome matrix for the control group is equal to one.

Theorem 2. $\Sigma_{i,j}(\mathbf{PO}^C) = 1$

Proof of Theorem 2

$$\begin{aligned} \Sigma_{i,j}(\mathbf{PO}^C) &= \pi_0^C F_0^C(t) + \pi_0^C\left(1-F_0^C(t)\right) + \pi_1^C F_1^C(t) + \pi_1^C\left(1-F_1^C(t)\right) \\ &= \pi_0^C F_0^C(t) - \pi_0^C F_0^C(t) + \pi_0^C + \pi_1^C F_1^C(t) - \pi_1^C F_1^C(t) + \pi_1^C \\ &= \pi_0^C + \pi_1^C \\ &= 1 \quad \square \end{aligned}$$

Likewise, the sum of the elements of the prescription-outcome matrix for the treatment group can be proven to be equal to one.

4.2. Performance measures

A number of measures that have been proposed in the literature for assessing the performance of causal classification models can be defined in terms of the prescription-outcome matrices of the control and treatment group. To this end, we first define the action rate, which we denote with $\dot{\phi}(t)$ or $\dot{\phi}$ in short, as the causal equivalent of the predicted positive rate, ϕ.

Definition 4. The action rate is the proportion of instances in $\mathcal{D} = \{\mathcal{D}^T, \mathcal{D}^C\}$ with a score below the threshold t:

$$\dot{\phi}(t) := \frac{\left(\pi_0^C F_0^C(t) + \pi_1^C F_1^C(t)\right) + \left(\pi_0^T F_0^T(t) + \pi_1^T F_1^T(t)\right)}{2}. \tag{23}$$

An action rate $\dot{\phi}(t)$ can be defined to causally classify instances, which corresponds to the use of a single threshold t for defining the action segments in both the control and treatment group. Note that the resulting action rates within the treatment group and control group, which we denote with $\dot{\phi}(t)^T$ and $\dot{\phi}(t)^C$ for $\mathcal{D}^T$ and $\mathcal{D}^C$, respectively, may be different from the overall action rate $\dot{\phi}(t)$ across both groups, i.e., in $\mathcal{D}$. As discussed in Devriendt et al. (2020), an alternative approach that is adopted in literature consists in using a different threshold for the control and treatment group, so as to achieve the same action rate in the control and treatment group separately.

The Qini curve (Radcliffe 2007), $\text{Qini}(\dot{\phi})$, is a frequently used visual approach for representing causal classification model performance. It is the causal equivalent of the gains curve for classification and reports the cumulative increase in positive outcomes as a function of the action rate $\dot{\phi}$:

$$\text{Qini}(\dot{\phi}) := \pi_0^T F_0^T(\dot{\phi}) - \pi_0^C F_0^C(\dot{\phi}). \tag{24}$$

Whereas the gains curve is a monotonically increasing function, the Qini curve may not be. Since the effect of the action $W = 1$ can be positive as well as negative (or null), the increase in positive outcomes for an action rate below 100% (i.e., $Q(\dot{\phi})$ for $\dot{\phi} < 1$.) may exceed the value that is obtained for an action rate equal to 100% (i.e., $Q(\dot{\phi})$ for $\dot{\phi} = 1$). In other words, the net increase in positive outcomes, which is the sum of both positive and negative effects, when applying the action to all subjects may be smaller than the net increase in positive outcomes when applying the action to a subset of subjects, e.g., for the subset of subjects with a positive effect excluding the subjects with no or a negative effect.

The Qini coefficient, Q, is the causal counterpart of the Gini coefficient (Radcliffe 2007). It is equal to the ratio of the area between the Qini curve and the diagonal, and the area between the Qini curve of a perfect model and the diagonal.

$$Q := \frac{\int \text{Qini}(\dot{\phi}) \mathrm{d}\dot{\phi} - \frac{\pi_0^T - \pi_0^C}{2}}{\frac{\pi_0^T - {\pi_0^C}^2 - {\pi_0^T}^2}{2}}. \tag{25}$$

Note that the diagonal represents the Qini curve of a random model. Hence, a random model achieves a Qini coefficient of zero. The perfect model is defined as achieving the maximum increase in positive outcomes, which is equal to the positive outcome rate in the treatment group, for an action rate that is equal to the positive outcome rate in the treatment group (Radcliffe 2007). When defining the Qini curve of the perfect model as such, it is assumed that all positive outcomes in the treatment group are caused by the action and would have been negative when the action were not applied, and that all positive instances in the control group would have been negative when the action were applied. These assumptions regarding the perfect model make sure that the

Qini coefficient scales in between zero and one, but are unrealistic and therefore negatively affect the interpretability of the Qini coefficient.

As an alternative approach, potential negative effects may be ignored in defining the Qini curve of the perfect model, which as a result is monotonically increasing. The resulting coefficient is not bounded above by one and is called the little Qini coefficient, q_0 (Radcliffe 2007):

$$\begin{aligned} q &:= \frac{\int \mathrm{Qini}(\dot{\phi})\mathrm{d}\dot{\phi} - \frac{\pi_0^T - \pi_0^C}{2}}{\frac{\pi_0^T - \pi_0^C}{2} - \frac{(\pi_0^T - \pi_0^C)^2}{2}}, \\ &= \frac{2\int \frac{\mathrm{Qini}(\dot{\phi})}{\pi_0^T - \pi_0^C}\mathrm{d}\dot{\phi} - 1}{1 - (\pi_0^T - \pi_0^C)}. \end{aligned} \tag{26}$$

Note that there is a theoretical upper bound to the little Qini coefficient, which can be obtained by using the Qini curve of the perfect model in the denominator of Equation 25 in integral in the numerator of Equation 26.

The liftup curve, $\dot{\lambda}(\dot{\phi})$, is the causal counterpart of the lift curve (Devriendt et al. 2019) and plots the ratio of the increase in positive outcomes for action rate $\dot{\phi}$ and the increase in positive outcomes for $\dot{\phi} = 1$, as a function of $\dot{\phi}$:

$$\begin{aligned} \dot{\lambda}(\dot{\phi}) &:= \frac{\pi_0^T F_0^T(\dot{\phi}) - \pi_0^C F_0^C(\dot{\phi})}{\dot{\phi}(\pi_0^T - \pi_0^C)}, \\ &= \frac{\mathrm{Qini}(\dot{\phi})}{(\pi_0^T - \pi_0^C)\dot{\phi}}. \end{aligned} \tag{27}$$

Whereas the Qini and liftup curve are the causal counterparts of the gains and lift curve, to our knowledge no causal counterpart for the ROC curve an the associated AUC measure has been proposed. Such a causal counterpart, however, is straightforward to define. To this end, we consider that, at some threshold t, positive outcomes in the action segment of the treatment group and negative outcomes in the action segment of the control group are the equivalent of true positives in a conventional classification context; on the other hand, negative outcomes in the action segment of the treatment group and positive outcomes in the action segment of the control group can be considered as the equivalent of false negatives. As such, we transform the causal classification problem into a conventional classification problem. This allows to evaluate a causal classification model using the ROC curve and the AUC measure. A similar approach for learning causal classification models was initially proposed by Lai et al. (2006)

Note that, a conceptually similar approach was adopted in Devriendt et al. (2020), who propose a novel measure, i.e., the promoted cumulative gain, for use as the objective function with learning to rank methods for causal classification. Note that such an approach may strike as perverse and missing the true objective, since hinting toward a straightforward solution that maximizes performance in terms of this measure by developing two classification models, a first model ranking treatment group observations from low to high probability to be positive and a second model ranking control group observations from low to high probability to be negative. However, when evaluating a causal classification model on a test set, it is to rank instances without knowing whether an instance pertains to the treatment or control group. Hence, the model that is used for ranking cannot be selected depending on the group membership of an instance and thus the conceived solution is invalid. An essential assumption underlying causal learning methods is that group membership cannot and should not be implicitly learned so as to maximize performance by using group membership information to classify. To this end, we either need the samples to be i.i.d., or, when they are not, they may be balanced using propensity scores or the effect of using imbalanced samples may be countered, e.g., by learning a balanced representation through domain adversarial training (Ganin et al. 2016).

We first define causal sensitivity and causal false discovery rate, which we denote with $\psi(t)$, as the causal counterparts of sensitivity and false discovery rate, respectively:

Definition 5.

$$\text{Causal sensitivity} := \frac{\pi_0^T F_0^T(t) + \pi_1^C F_1^C(t)}{\pi_0^T + \pi_1^C} \tag{28}$$

$$\psi(t) := \frac{\pi_1^T F_1^T(t) + \pi_0^C F_0^C(t)}{\pi_1^T + \pi_0^C} \tag{29}$$

This allows to define the V-curve as the causal equivalent of the ROC curve, plotting the causal sensitivity in function of the causal false discovery rate, $\psi(t)$:

Definition 6.

$$V(\psi) := \frac{\pi_0^T F_0^T(\psi) + \pi_1^C F_1^C(\psi)}{\pi_0^T + \pi_1^C} \tag{30}$$

We define the V-measure, equal to the area under the V-curve, as the causal equivalent of the AUC measure:

Definition 7.

$$V := \int \left(\frac{\pi_0^T F_0^T(s) + \pi_1^C F_1^C(s)}{\pi_0^T + \pi_1^C} \right) \left(\frac{\pi_1^T f_1^T(s) + \pi_0^C f_0^C(s)}{\pi_1^T + \pi_0^C} \right) \mathrm{d}s \tag{31}$$

Theorem 3. *Classification, as defined in terms of the AUC, Gini and lift measures, is a specific case of causal classification as defined in terms of the V, little Qini and liftup measures, with the number of actions that can be applied exactly equal to one, i.e.,* $|W| = 1$.

Proof of Theorem 3 If the number of actions that can be applied is exactly equal to one, $W \in \{1\}$, then all instances in the data set for learning a causal classification model concern instances of the treatment group.

Hence, we have that $\pi_0^C = \pi_1^C = 0$.

It is straightforward to proof that, by setting $\pi_0^C = \pi_1^C = 0$ in Equations 25, 27 and 31 of the little Qini, liftup and V-measure, respectively, we obtain the Gini, lift and AUC measures as defined by Equations 7, 8 and 5, respectively.

Theorem 3 means that, from a mathematical perspective, (the evaluation of) a conventional classification model is a specific case of (the evaluation of) a causal classification problem. The evaluation of a causal classification model typically is performed by comparing the outcomes in the treatment sample with the outcomes in the control sample. The latter is non-existent in a conventional classification context, where we evaluate performance in absolute sense, at least in terms of the effect of the single action that can be applied, whereas in causal classification we evaluate performance in relative sense, i.e., by comparing to the baseline of the control group. Note that, when only a treatment sample is available, all instances have received the same treatment and the interest is in identifying instances pertaining to the positive class.

In the following sections, we generalize the profit formula for evaluating a conventional classification model, as introduced in Equation 10, for causal classification.

5. Cost-sensitive causal classification performance

In this section, we further extend upon the evaluation framework that has been developed in the previous sections and introduce cost-sensitive performance measures for the evaluation of causal classification models.

5.1. Profit

A cost-benefit causal classification framework is needed to allow the evaluation of causal classification models based on the profit that results from applying the model. The need for such a framework stems from its potential use (1) for evaluating the performance of causal classification models, but as well (2) for defining an objective function and learning a causal classification model from data in line with the business objective.

The prescription-outcome matrices report the observed outcome distributions in the treatment and control groups for the action and no-action segments. Note that the outcomes in the no-action segment of the treatment group concern, per definition, outcomes that are observed for subjects that were exposed to the action. Vice versa, the outcomes in the action-segment of the control group are observed for subjects that were not exposed to the action. For reporting the outcome distribution that would be observed for the action and no-action segment at threshold t, i.e., when the prescribed action would have been applied, we introduce the action-outcome matrix.

Definition 8. The action-outcome matrix, **AO**, indicates the proportion of positive and negative outcomes in the action and no-action segment if the prescribed action would effectively be applied to the population as segmented, i.e., causally classified, by a causal classification model for threshold t.

$$\mathbf{AO} := \begin{array}{c} \text{Applied action} \\ \begin{array}{cc} W=1 & W=0 \end{array} \\ \begin{bmatrix} \pi_0^T F_0^T(t) & \pi_0^T\big(1-F_0^C(t)\big) \\ \pi_1^T F_1^T(t) & \pi_1^T\big(1-F_1^C(t)\big) \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \text{Y=0} \\ \text{Y=1} \end{array} \quad \text{Outcome} \tag{32}$$

For the action segment, the outcome distribution is obtained from the treatment group prescription-outcome matrix for threshold t. The outcome distribution for the no-action segment, on the other hand, is obtained from the control group prescription-outcome matrix for the same threshold t. Note that the action-outcome matrix is equivalent with the treatment-response matrix as introduced by Kane et al. (2014).

We proof that the sum of the elements of the action-outcome matrix is equal to one.

Theorem 4. $\Sigma_{i,j}(\mathbf{AO}) = 1$

Proof of Theorem 4

$$\begin{aligned} \Sigma_{i,j}(\mathbf{AO}) &= \pi_0^T F_0^T(t) + \pi_0^C\big(1-F_0^C(t)\big) + \pi_1^T F_T^C(t) + \pi_1^C\big(1-F_1^C(t)\big) \\ &= \pi_0^T F_0^T(t) + \pi_0^C - \pi_0^C F_0^C(t) + \pi_1^T F_T^C(t) + \pi_1^C - \pi_1^C F_1^T(t) \\ &= 1 + \pi_0^T F_0^T(t) + \pi_1^T F_1^T(t) - \pi_1^C F_1^C(t) - \pi_0^T F_0^T(t) \end{aligned}$$

Hence, we need to proof that:

$$\pi_0^T F_0^T(t) + \pi_1^T F_1^T(t) - \pi_1^C F_1^C(t) - \pi_0^T F_0^T(t) = 0$$

Or:

$$\pi_0^T F_0^T(t) + \pi_1^T F_1^T(t) = \pi_1^C F_1^C(t) + \pi_0^T F_0^T(t)$$

We define the cumulative distribution of treatment and control group instances in function of the causal model score s, which we denote with $F^T(s)$ and $F^C(s)$, respectively:

$$\begin{aligned} F^T(t) &:= \pi_0^T F_0^C(t) + \pi_1^T F_1^T(t) \\ F^C(t) &:= \pi_0^C F_0^C(t) + \pi_1^C F_1^C(t) \end{aligned}$$

Hence, $\Sigma_{i,j}(\mathbf{AO}) = 1$ if $F^T(t) = F^C(t)$, which holds by definition, since treatment and control group are identically distributed random samples from the population. $\square$

The profit of adopting a causal classification model is determined by the outcomes and the actions that are applied, as reported in the action-outcome matrix. For calculating the profit, we introduce two matrices:

DEFINITION 9. The outcome-benefit matrix, **OB**, indicates the benefit of an outcome depending on the action that is applied. The benefit of obtaining an outcome from class i after applying action j is denoted with $b_{(i,j)}$.

$$\mathbf{OB} := \begin{array}{cc} \multicolumn{2}{c}{\text{Applied action}} \\ W=1 & W=0 \\ \left[\begin{array}{cc} b_{(0,1)} & b_{(0,0)} \\ b_{(1,1)} & b_{(1,0)} \end{array}\right] & \end{array} \begin{array}{c} \\ \\ \text{Y=0} \\ \text{Y=1} \end{array} \ \text{Outcome} \tag{33}$$

DEFINITION 10. The action-cost matrix, **AC**, indicates the cost of an action depending on the outcome. The cost of applying action i when obtaining an outcome from class j is denoted with $c_{(i,j)}$.

$$\mathbf{AC} := \begin{array}{cc} \multicolumn{2}{c}{\text{Applied action}} \\ W=1 & W=0 \\ \left[\begin{array}{cc} c_{(0,1)} & c_{(0,0)} \\ c_{(1,1)} & c_{(1,0)} \end{array}\right] & \end{array} \begin{array}{c} \\ \\ \text{Y=0} \\ \text{Y=1} \end{array} \ \text{Outcome} \tag{34}$$

By convention, the benefit and cost values in the outcome-benefit and action-cost matrix are positive or equal to zero.

Then, the profit per instance, $\dot{P}$, of adopting a causal classification model $\dot{m}$ at threshold t can be calculated as the difference between the benefit per instance resulting from the obtained outcomes and the cost per instance resulting from the applied actions:

$$\begin{aligned}
\dot{P} &= \Sigma_{i,j}(\mathbf{AO} \circ \mathbf{OB} - \mathbf{AO} \circ \mathbf{AC}), \\
&= \Sigma_{i,j}\big(\mathbf{AO} \circ (\mathbf{OB} - \mathbf{AC})\big), \\
&= \Sigma_{i,j}(\mathbf{AO} \circ \dot{\mathbf{C}}\mathbf{B}) \qquad (35)
\end{aligned}$$

DEFINITION 11. The causal cost-benefit matrix, $\dot{\mathbf{C}}\mathbf{B}$, indicates the profit of an outcome depending on the action that is applied and can be calculated from the outcome-benefit and action-cost matrix as follows:

$$\dot{\mathbf{C}}\mathbf{B} = \mathbf{OB} - \mathbf{AC}. \tag{36}$$

Values in the $\dot{\mathbf{CB}}$ matrix are positive for action-outcome combinations which yield a net profit, and values will be negative for combination which yield a net loss.

Note the strong resemblance between Equation 10 and Equation 35, with the action-outcome matrix being the causal counterpart of the confusion matrix and the causal cost-benefit matrix being the causal counterpart of the cost-benefit matrix. Then, following the notation introduced above, we obtain the following expression for the profit per instance:

$$\begin{aligned} P = {} & \pi_0^T F_0^T(t)(b_{(0,1)} - c_{(0,1)}) + \pi_0^C\big(1 - F_0^C(t)\big)(b_{(0,0)} - c_{(0,0)}), \\ & + \pi_1^T F_1^T(t)(b_{(1,0)} - c_{(1,0)}) + \pi_1^C\big(1 - F_1^C(t)\big)(b_{(1,1)} - c_{(1,1)}). \end{aligned} \tag{37}$$

5.2. Causal profit

The formula of Equation 35 for calculating the profit implicitly establishes a baseline which defines the meaning of the value zero, i.e., no actions are applied and no outcomes are obtained. As such, the resulting profit is the absolute profit, as defined in Section 3.

In a causal classification problem context, however, this is not a meaningful baseline scenario, since in practical applications an outcome will always occur for the subjects in the population at hand. A particular action may be applied to subjects that may change that outcome. In fact, a causal classification model is developed exactly for the purpose of optimizing the selection of subjects that are to be exposed to the action, with the eventual aim in a business context to increase profitability. In such a problem setting, the profit is to be calculated relative to the baseline scenario of not applying the action, to none of the subjects, which we henceforth call the no-action baseline scenario.

In the no-action baseline scenario, the no-action segment comprises the entire population and the baseline outcome distribution is reported by the baseline action-outcome matrix, $\mathbf{AO}_b$:

$$\mathbf{AO}_b = \begin{bmatrix} 0 & \pi_0^C \\ 0 & \pi_1^C \end{bmatrix}. \tag{38}$$

Note that, by definition, $\mathbf{AO}_b = \mathbf{PO}^C$ for $t = max(s)$.

Definition 12. We define the causal effect matrix, $\dot{\mathbf{E}}$, reporting the shift in the outcome distribution that results from applying the action $W = 1$ to the action-segment. The causal effect matrix is calculated as the difference between the action-outcome matrix and the action-outcome matrix of the no-action baseline scenario, which we denote with $\mathbf{AO}_b$.

$$\dot{\mathbf{E}} = \mathbf{AO} - \mathbf{AO}_b \tag{39}$$

The causal effect matrix is the causal equivalent of the effect matrix as defined in Equation 15.

We find the causal effect matrix in Equation 40 by elaborating Equation 39 using the action-outcome matrices, **AO** and $\mathbf{AO}_b$ as defined in Equation 32 and Equation 38, respectively, and with $\pi_0^C\left(1-F_0^C(t)\right)-\pi_0^C=-\pi_0^C F_0^C(t)$ and $\pi_1^C\left(1-F_1^C(t)\right)-\pi_1^C=-\pi_1^C F_1^C(t)$:

$$\dot{\mathbf{E}} := \begin{array}{c} \text{Applied action} \\ \begin{array}{cc} W=1 & W=0 \end{array} \\ \left[\begin{array}{cc} \pi_0^T F_0^T(t) & -F_0^C(t) \\ \pi_1^T F_1^T(t) & -F_1^C(t) \end{array}\right] \end{array} \begin{array}{c} \text{Y=0} \\ \text{Y=1} \end{array} \text{Outcome} \tag{40}$$

The causal effect matrix reports the shift in the outcome distribution of both the action and no-action segment, caused by applying the action to the action-segment and compared to the no-action baseline scenario. Intuitively, we expect that the aggregated decrease in the number of instances of the positive and negative class in the no-action segment is equal to the aggregated increase in the number of instances of the positive and negative class in the action-segment. We proof that the sum of the elements of the causal effect matrix is equal to zero.

THEOREM 5. $\Sigma_{i,j}(\dot{\mathbf{E}})=0$

Proof of Theorem 5 Following Theorem 4, we have:

$$\Sigma_{i,j}(\mathbf{AO})=1$$

From Equation 38, we obtain:

$$\Sigma_{i,j}(\mathbf{AO}_b)=\pi_0^C+\pi_1^C=1$$

Hence:

$$\begin{aligned} \Sigma_{i,j}(\dot{\mathbf{E}}) =& \Sigma_{i,j}(\mathbf{AO}-\mathbf{AO}_b) \\ =& \Sigma_{i,j}(\mathbf{AO})-\Sigma_{i,j}(\mathbf{AO}_b) \\ =& 1-1 \\ =& 0 \quad \square \end{aligned}$$

DEFINITION 13. We define the causal profit per instance, which we denote $\dot{P}$, as the difference between the profit per instance, P, as defined in Equation 35, and the profit per instance that is obtained in the no-action baseline scenario, which we denote P_b:

$$\dot{P} := P - P_b \tag{41}$$

We proof the following theorem, which establishes the relation between the causal profit, the causal effect matrix and the profit matrix:

Theorem 6. $\dot{P} = \Sigma_{i,j}(\dot{\mathbf{E}} \circ \dot{\mathbf{CB}})$

Proof of Theorem 6

$$
\begin{aligned}
\dot{P} &= P - P_b \\
&= \Sigma_{i,j}(\mathbf{AO} \circ \mathbf{OB} - \mathbf{AO} \circ \mathbf{AC}) - \Sigma_{i,j}(\mathbf{AO}_b \circ \mathbf{OB} - \mathbf{AO}_b \circ \mathbf{AC}) \\
&= \Sigma_{i,j}\big((\mathbf{AO} \circ \mathbf{OB} - \mathbf{AO} \circ \mathbf{AC}) - (\mathbf{AO}_b \circ \mathbf{OB} - \mathbf{AO}_b \circ \mathbf{AC})\big) \\
&= \Sigma_{i,j}\big((\mathbf{AO} - \mathbf{AO}_b) \circ \mathbf{OB} - (\mathbf{AO} - \mathbf{AO}_b) \circ \mathbf{AC}\big) \\
&= \Sigma_{i,j}(\dot{\mathbf{E}} \circ \mathbf{OB} - \dot{\mathbf{E}} \circ \mathbf{AC}) \\
&= \Sigma_{i,j}\big(\dot{\mathbf{E}} \circ (\mathbf{OB} - \mathbf{AC})\big) \\
&= \Sigma_{i,j}(\dot{\mathbf{E}} \circ \dot{\mathbf{CB}}) \quad \square
\end{aligned}
$$

By elaborating Equation 41 using Theorem 6, we obtain the following formula to calculate the causal profit, $\dot{P}$:

$$
\begin{aligned}
\dot{P} = &\pi_0^T F_0^T(t)(b_{(0,1)} - c_{(0,1)}) - \pi_0^C F_0^C(t)(b_{(0,0)} - c_{(0,0)}) \\
&+ \pi_1^T F_1^T(t)(b_{(1,1)} - c_{(1,1)}) - \pi_1^C F_1^C(t)(b_{(1,0)} - c_{(1,0)}).
\end{aligned} \tag{42}
$$

The definition of causal profit in Equation 41 is equivalent with the definition of relative profit in Equation 13 for conventional classification. Whereas the performance of a conventional classification model is typically evaluated in a relative manner by comparing to a baseline model, as discussed in Section 3, in causal classification the baseline is a model which prescribes a baseline action to be applied to all instances, e.g., no action.

5.3. Maximum Causal Profit and Expected Maximum Causal Profit

The profit as calculated following Equation 42 is a function of the threshold t. Extending upon the Maximum Profit measure for classification, we propose to maximize the causal profit per instance as expressed by Equation 42 by optimizing the threshold t. The resulting maximum causal profit per instance, $\dot{MP}$, can be used as a measure for evaluating the performance of a causal classification model.

Definition 14. The Maximum Causal Profit measure for evaluating a causal classification model is defined as follows::

$$
\dot{MP} := \max_{\forall t} \big(\dot{P}(t; \mathbf{OB}, \mathbf{AC})\big) \tag{43}
$$

$$= \max_{\forall t} \left(\Sigma_{i,j} (\dot{\mathbf{E}} \circ \dot{\mathbf{CB}}) \right)$$
$$= \max_{\forall t} \left(\Sigma_{i,j} \left(\dot{\mathbf{E}} \circ (\mathbf{OB} - \mathbf{AC}) \right) \right)$$
$$= \dot{P}(t^*; \mathbf{OB}, \mathbf{AC})$$

Definition 15. The profit maximizing fraction η_{MCP} is the proportion of the population to which the action should be applied in order to maximize the profit of the intervention:

$$\eta_{\dot{M}P} = \pi_0^T F_0^T(t^*) + \pi_1^T F_1^T(t^*) \tag{44}$$

The maximum causal profit measure is an intuitively interpretable measure that allows evaluation and comparison of causal classification models in terms of its potential for maximizing the business outcome. The $\dot{M}P$ measure may complement the use of cost-insensitive performance measures, such as the Qini, liftup or V -measure. Moreover, it supports practitioners in decision-making by providing the optimal threshold, t^*, and the associated and profit maximizing action rate for segmenting the population into an action and no-action segment based on the causal classification model score.

In a similar manner as the MP measure for classification was extended to the EMP measure, we extend upon the $\dot{M}P$ measure and define the Expected Maximum Causal Profit measure.

Definition 16. The Expected Maximum Causal Profit ($E\dot{M}P$) measure for evaluating a causal classification model is defined as follows:

$$E\dot{M}P := \int_{\mathbf{OB}} \int_{\mathbf{AC}} \dot{P}(t^*, \mathbf{OB}, \mathbf{AC}) \cdot h(\mathbf{OB}, \mathbf{AC}) d\mathbf{AC} d\mathbf{OB} \tag{45}$$

with $h(\mathbf{OB}, \mathbf{AC})$ the joint distribution of the cost and benefit parameters in the action-cost and outcome-benefit matrices.

Note that in practical applications, many of the cost and benefit parameters may take a value of zero and a limited number of non-zero parameters may be uncertain or random, which facilitates the practical calculation of the $E\dot{M}P$ measure.

The main advantages of the $E\dot{M}P$ measure are its robustness and the flexibility it provides to the user in adapting to the applicable operating conditions. Further extensions are possible in line with Hernández-Orallo et al. (2012) to address the uncertainty regarding the operating conditions at time of model evaluation. The main disadvantages of the $E\dot{M}P$ measure are the difficulty in specifying an appropriate joint probability distribution for the cost-benefit parameters and the added complexity. The interpretation of the measure remains the same as for the $\dot{M}P$ measure. Whereas the $E\dot{M}P$ measures relies upon the representativeness of the sample and the ability of the user to specify an accurate value for the cost and benefit parameters, the $E\dot{M}P$ measure provides the means to achieve a measure which better reflects the true average profit that will be obtained when applying the causal classification model under a range of operating conditions.

When comparing on the one hand side Equations and of the $\dot{M}P$ and $E\dot{M}P$, respectively, and on the other hand side Equations and of the MP and EMP, respectively, we find that the causal counterpart of the confusion matrix is the action-outcome matrix, of the effect matrix the causal effect matrix, and of the cost-benefit matrix the causal cost-benefit matrix. Moreover, extending upon Theorem 3, we can proof that that conventional classification, from a cost-sensitive evaluation perspective, is a specific case of causal classification.

THEOREM 7. *Classification as defined in terms of the profit, maximum profit and expected maximum profit measures, is a specific case of causal classification as defined in terms of the causal profit, maximum causal profit and expected maximum causal profit measures, with the number of actions that can be applied exactly equal to one, i.e.,* $|W| = 1$.

Proof of Theorem 7 If the number of actions that can be applied is exactly equal to one, $W \in \{1\}$, then action $W = 1$ (which may represent any baseline action, including no action) is prescribed for both the instances in the action and no-action segments, as defined by the decision threshold t.
As a result, the action-outcome matrix, **AO**, as defined in Equation 32, instantiates to a confusion matrix, **CF**, as defined in Equation 1.
Hence, the interpretation and formula for calculating causal profit, $\dot{P}$ as defined by Equation 35, is identical to the interpretation and formula for calculating classification profit, P, as defined in Equation 10, with the causal cost-benefit matrix, $\dot{\mathbf{CB}}$, as defined in Equation 36 instantiating to the classification cost-benefit matrix, **CB**, as defined in Equation 9.

6. Practical applications

In this section, we instantiate the causal profit measure to evaluate causal classifiers for customer retention and customer response modeling, as previously proposed in literature.

6.1. Customer retention

Devriendt et al. (2019) apply causal classification methods for developing customer churn uplift models, allowing to estimate the effect of a retention campaign on the churn risk of an individual customer. This allows to optimize the selection of customers that are to be targeted with the retention campaign. Devriendt et al. (2019) propose the following formula to calculate the total profit of a retention campaign, $P^{\text{Retention}}$ (Equation 12, page 18):

$$P^{\text{Retention}} = N\alpha\big((\beta_C - \beta_T)(CLV - C_c - C_i) - (1 - \beta_C)(C_c + C_i) - \beta_T C_c\big), \tag{46}$$

with N the total number of customers and α the proportion of customers that are targeted; hence, $N\alpha$ is the number of customers in the action segment; β_C and β_T are the churn rates in the control and treatment group action segment, respectively; CLV is the customer lifetime value, i.e., the benefit of a customer that does not churn; C_c is the cost of contacting a customer and C_i is the cost of the incentive that is offered in the retention campaign.

Equation 46 can be instantiated from the generic causal profit formula of Equation 41. The maximum profit uplift (MPU) measure, as introduced in Devriendt et al. (2019) and which maximizes the profit calculated following Equation 46, hence is an instantiation of the general $\dot{MP}$ measure. The outcome of interest is whether a customer will soon churn ($Y = 1$) or not ($Y = 0$). The decision that is to be made is whether to target a customer with a retention campaign ($W = 1$) or not ($W = 0$), offering an incentive to remain loyal, e.g., a discount. The retention campaign is the action that may alter the outcome. For a detailed discussion on customer churn uplift modeling, one may refer to Ascarza (2018).

Table 1a shows the applicable outcome-benefit matrix. The benefit of a customer that does not churn, i.e., the benefit of a positive outcome, is independent of the action that is applied and equal to the customer lifetime value. The CLV is the average net discounted cash flow that is expected to be generated by a customer over a predefined time window. Table 1b shows the applicable action-cost matrix. Targeting a customer with a retention campaign involves a cost of contacting the customer, c_c. If the customer accepts the retention offer, an additional cost is incurred, i.e., the cost of the incentive, c_i. Remark that the cost of the incentive is only incurred when the customer does not churn, i.e., for a positive outcome. In this application, no-action means not targeting a customer, which does not involve a cost. The outcome-benefit and action-cost matrices align with the benefits and costs in the profit formula proposed in (Neslin et al. 2006), which underlies both the maximum profit measure for churn (Verbeke et al. 2012) and the MPU measure.

Given the outcome-benefit and action-cost matrix of Table 1, we can elaborate Equation 42 for customer retention as follows:

$$\dot{P} = \pi_0^T F_0^T(t)(CLV - C_c - C_i) - \pi_0^C F_0^C(t) CLV - \pi_1^T F_1^T(t) C_c. \tag{47}$$

Note that β_T and β_C in Equation 46 denote the churn rate in the action-segment of the treatment and the control group, respectively, and $(1 - \beta_T)$ and $(1 - \beta_C)$ indicate the proportion of non-churners in the action-segment of the treatment and the control group, respectively.

Table 1 Outcome-benefit and action-cost matrices for customer churn uplift modeling.

(a) OB matrix

Outcome	Applied action: W=1	W=0
Y=0	CLV	CLV
Y=1	0	0

(b) AC matrix

Outcome	Applied action: W=1	W=0
Y=0	C_c+C_i	0
Y=1	C_c	0

Using the following equalities:

$$\begin{aligned}
\pi_0^T F_0^T(t) &= \alpha(1-\beta_T), \\
&= \alpha(1-\beta_T+\beta_C-\beta_C), \\
&= \alpha(1-\beta_C)+\alpha(\beta_C-\beta_T), && (48) \\
\pi_0^C F_0^C(t) &= \alpha(1-\beta_C), && (49) \\
\pi_1^T F_1^T(t) &= \alpha\beta_T, && (50)
\end{aligned}$$

we can rework the causal profit per instance formula of Equation 47 to arrive at Equation 46:

$$\begin{aligned}
\dot{P} &= \big(\alpha(1-\beta_C)+\alpha(\beta_C-\beta_T)\big)(CLV-C_c-C_i)-\alpha(1-\beta_C)CLV-\alpha\beta_T C_c, \\
&= \alpha(\beta_C-\beta_T)\big)(CLV-C_c-C_i)+\alpha(1-\beta_C)(CLV-C_c-C_i)-\alpha(1-\beta_C)CLV-\alpha\beta_T C_c, \\
&= \alpha(\beta_C-\beta_T)\big)(CLV-C_c-C_i)+\alpha(1-\beta_C)CLV-\alpha(1-\beta_C)(C_c+C_i)-\alpha(1-\beta_C)CLV-\alpha\beta_T C_c, \\
&= \alpha(\beta_C-\beta_T)\big)(CLV-C_c-C_i)-\alpha(1-\beta_C)(C_c+C_i)-\alpha\beta_T C_c, \\
&= \frac{P^{\text{Retention}}}{N} && (51)
\end{aligned}$$

6.2. Customer response

Gubela et al. (2019) apply causal classification methods for developing customer response models, allowing to estimate the incremental effect of a discount on the propensity to purchase of an individual customer. This allows to optimize the selection of customers that are to be targeted with the marketing campaign. Gubela et al. (2019) propose the following formula to calculate the total profit of a response campaign, P^{Response} (Equation 13, page 653):

$$P^{\text{Response}} = N_\tau(\pi_\tau\delta_\tau-\pi_\zeta\delta_\zeta)-N_\tau\varepsilon_{unit}-\rho\sum_{i=1}^{N_\tau}\delta_{\pi_+,i}, \tag{52}$$

with N_τ the number of instances in the action-segment, with N equal to the total number of customers and τ the proportion of customers that are targeted; π_τ is the proportion of positive outcomes in the action-segment, and δ_τ is the average revenue in the action-segment; the equivalent quantities as measured for a control group, are π_ζ, and δ_ζ, respectively; ε_{unit} is the cost of targeting a customer with the marketing campaign; ρ is the percentage discount that is given, π_+ is the proportion of responders in the action-segment, and $\delta_{\pi_+,i}$ the revenue generated by customer i in the action-segment.

Equation 52 can be instantiated from the generic causal profit formula of Equation **??**. The outcome of interest is whether a customer makes a purchase ($Y=0$) or not ($Y=1$). The decision to be made is whether to target a customer with a response campaign ($W=1$) or not ($W=0$), e.g., offering a discount on a purchase. The response campaign is the action that may alter the outcome.

Table 2 Outcome-benefit and action-cost matrices for customer response uplift modeling.

		Applied action	
		W=1	W=0
Outcome	Y=0	δ^T	δ^C
	Y=1	0	0

(a) OB matrix

		Applied action	
		W=1	W=0
Outcome	Y=0	C_c+C_i	0
	Y=1	C_c	0

(b) AC matrix

Table 2 shows the outcome-benefit matrix and the action-cost matrix for this case. Note that the benefit of a positive outcome, here, depends on the action. We denote the constant cost of targeting a customer with $C_c=\varepsilon_{unit}$, and the average cost of the incentive with C_i. Gubela et al. (2019) consider a varying incentive cost. That is, the financial value of the promotional action varies according to the size of the discount and the shopping basket. In Equation **??** for calculating the total profit, the total cost of the incentives is calculated. In order to calculate the profit per instance, we will take the average cost of the incentive into account, which is calculated as the ratio of the total cost of the incentives and the number of responders in the action-segment, $N_\tau\pi_\tau$:

$$C_i=\frac{\rho\sum_{i=1}^{N_\tau}\delta_{\pi_+,i}}{N_\tau\pi_\tau} \tag{53}$$

Using the following equalities,

$$\begin{aligned}
\tau\pi_\tau &= \pi_0^T F_0^T(t),\\
\tau\pi_\zeta &= \pi_0^C F_0^C(t),\\
\tau &= \pi_0^T F_0^T(t)+\pi_1^T F_1^T(t),\\
\rho\sum_{i=1}^{N_\tau}\delta_{\pi_+,i} &= N\tau\pi_\tau C_i,
\end{aligned}$$

$$= N\pi_0^T F_0^T(t) C_i,$$

we can rework Equation 52 to arrive at the causal profit per instance formula of Equation 41:

$$\begin{aligned}
\frac{P^{\text{Response}}}{N} =& \tau\pi_\tau\delta_\tau - \tau\pi_\zeta\delta_\zeta - \tau\varepsilon_{unit} - \frac{\rho\sum_{i=1}^{N_\tau}\delta_{\pi_+,i}}{N},\\
=& \pi_0^T F_0^T(t)\delta_\tau - \pi_0^C F_0^C(t)\delta_\zeta - \pi_0^T F_0^T(t)C_c - \pi_1^T F_1^T(t)C_c - \pi_0^T F_0^T(t)C_i,\\
=& \pi_0^T F_0^T(t)(\delta_\tau - C_c - C_i) - \pi_0^C F_0^C(t)\delta_\zeta - \pi_1^T F_1^T(t)C_c,\\
=& \Sigma_{i,j}\big(\dot{\mathbf{E}} \circ (\mathbf{OB} - \mathbf{AC})\big),\\
=& \Sigma_{i,j}(\dot{\mathbf{E}} \circ \dot{\mathbf{C}}\mathbf{B}),\\
=& \dot{P}
\end{aligned}$$

7. Conclusions and research agenda

Classification models are adopted across a variety of operational business processes to predict the unknown outcome of cases in function of known characteristics. This allows to anticipate and optimize managerial decision-making, by taking actions based on the predictions of the model. In order to improve upon the performance of conventional classification models, cost-sensitive classification models can be adopted to take into account the cost and benefit of the various possible outcomes, or causal classification models can be adopted to take into account the causal effect of an action that is taken on the outcome of interest. The objective in this paper is to merge cost-sensitive and causal classification so as to practically support optimization of resource allocation and maximize intervention effectiveness (Gupta et al. 2020).

To this end, we introduce a unifying framework which allows to evaluate both causal and conventional classification models, in both a cost-sensitive as well as cost-insensitive manner. The framework's key constituents are a number of matrices that are introduced and defined in this paper, i.e., the prescription-outcome matrix (**PO**), the action-outcome matrix (**AO**), the action-cost matrix (**AC**), the outcome-benefit matrix (**OB**), and the causal cost-benefit matrix ($\dot{\mathbf{C}}\mathbf{B}$).

In developing the framework, we have formalized the use of various baseline models for evaluating classification model performance in a relative manner. The use of a baseline model is shown to be inherent to causal classification, but applies as well in conventional classification. We have formalized the choice for a baseline model by defining the effect (**E**) and causal effect matrix ($\dot{\mathbf{E}}$), which is equal to the difference between the confusion matrix of the model and of the baseline model in conventional classification, and between the action-outcome matrix of the model and the baseline model in causal classification, respectively.

The proposed framework is shown to instantiate to a range of existing, cost-insensitive performance measures for evaluating causal classification models, i.e., Qini, little Qini and liftup. Based on the framework, we define a novel visual approach and associated performance measure for causal classification model evaluation, i.e., the V-curve and V-measure, which are the causal counterparts of the ROC curve and the AUC measure for conventional classification, respectively.

Based on the presented framework, we propose three novel cost-sensitive measures for evaluating causal classification models, i.e., profit (or equivalently, loss), maximum profit and expected maximum profit. These measures take into account both the causal effect and the cost of an action, as well as the cost or benefit of an outcome, in assessing the performance of a causal classification model. This allows to optimize the threshold that is to be adopted for deciding in which cases the envisioned action is to be applied, with the objective of maximizing the profit that results from applying the action in the selected cases and not applying the action in the non-selected cases, based on the predicted causal effect of the action on the outcome by the causal classification model. As such, we effectively merge cost-sensitive and causal classification and present a practical approach that supports business decision-making.

In addition, based on the framework, we proof that a range of causal classification performance measures instantiate to a range of equivalent performance measures for assessing the performance of conventional classification model. Hence, we proof that conventional classification can be regarded as a specific case of causal classification, with the number of possible actions that can be applied equal to one. As such, the proposed framework unifies cost-sensitive, causal and conventional classification.

The proposed framework opens a variety of directions for future research. A first stream of topics concerns the extension of the proposed cost-sensitive causal classification framework beyond the binary action and binary outcome setting which was elaborated in this paper. Topics of prime interest in this stream concern an extension of the framework to multi-action or multi-treatment effect estimation, to multi-valued outcomes (similar to multi-class classification), to multiple outcomes, to continuous actions, to continuous outcomes, and to instance-dependent (also called observation-dependent or example-dependent) costs and benefits.

Another extension of interest concerns the cost of acquiring data through field experiments, which could be taken into account in adopting a broader perspective towards the use of causal classification (Simester et al. 2020).

A second stream involves the development of causal learning methods which adopt a cost-sensitive objective function and as such align with the presented cost-sensitive performance measures that have been introduced in this paper and as such allow to learn causal models which better align with their use for business decision-making. As a source of inspiration for developing cost-sensitive causal learning methods, we identify two possible perspectives in implementing a cost-sensitive objective function: (1) the cost-sensitive objective function may assess the overall ranking, across all possible thresholds that can be adopted, as usually adopted in cost-sensitive learning; (2) the cost-sensitive objective function may reflect the need for optimizing the threshold and maximizing the profitability, as adopted in profit-driven learners which aim to learning classifiers by optimizing maximum profit (Höppner et al. 2018, Stripling et al. 2018). Cost-sensitive causal learning methods may also be of practical use for improving performance and stability of causal classification models when the class distribution is imbalanced, since many cost-sensitive conventional classification methods have been developed with this objective in mind and shown to be effective in addressing this issue, or when either or both the treatment or control groups are small. These potential uses add to the importance of developing such methods.

Acknowledgments

The authors gratefully acknowledge the support of INNOVIRIS.

References

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. Journal of Marketing Research 55(1):80–98.

Ben Rhouma T, Zaccour G (2018) Optimal marketing strategies for the acquisition and retention of service subscribers. Management Science 64(6):2609–2627.

Devriendt F, Berrevoets J, Verbeke W (2019) Why you should stop predicting customer churn and start using uplift models. Information Sciences .

Devriendt F, Guns T, Verbeke W (2020) Learning to rank for uplift modeling. arXiv preprint arXiv:2002.05897 .

Devriendt F, Moldovan D, 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.

Elkan C (2001) The foundations of cost-sensitive learning. International Joint Conference on Artificial Intelligence, volume 17, 973–978 (Lawrence Erlbaum Associates Ltd).

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V (2016) Domain-adversarial training of neural networks. The Journal of Machine Learning Research 17(1):2096–2030.

Garrido F, Verbeke W, Bravo C (2018) A robust profit measure for binary classification model evaluation. Expert Systems with Applications 92:154–160.

Gubela RM, Lessmann S, Jaroszewicz S (2019) Response transformation and profit decomposition for revenue uplift modeling. European Journal of Operational Research .

Gupta V, Han BR, Kim SH, Paek H (2020) Maximizing intervention effectiveness. Management Science .

Hand DJ (2009) Measuring classifier performance: a coherent alternative to the area under the roc curve. Machine Learning 77(1):103–123.

Hernández-Orallo J, Flach P, Ferri C (2012) A unified view of performance metrics: translating threshold choice into expected classification loss. Journal of Machine Learning Research 13(Oct):2813–2869.

Holland PW (1986) Statistics and causal inference. Journal of the American Statistical Association 81(396):945–960.

Höppner S, Stripling E, Baesens B, vanden Broucke S, Verdonck T (2018) Profit driven decision trees for churn prediction. European Journal of Operational Research .

Imai K, Ratkovic M, et al. (2013) Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics 7(1):443–470.

Kane K, Lo VS, Zheng J (2014) Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. Journal of Marketing Analytics 2(4):218–238.

Lai YT, Wang K, Ling D, Shi H, Zhang J (2006) Direct marketing when there are voluntary buyers. 6th International Conference on Data Mining, ICDM2006, 922–927 (IEEE).

Neslin SA, Gupta S, Kamakura W, Lu J, Mason CH (2006) Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research 43(2):204–211.

Qian M, Murphy SA (2011) Performance guarantees for individualized treatment rules. Annals of Statistics 39(2):1180.

Radcliffe NJ (2007) Using control groups to target on predicted lift: Building and assessing uplift model.

Shalit U, Johansson FD, Sontag D (2017) Estimating individual treatment effect: generalization bounds and algorithms. Proceedings of the 34th International Conference on Machine Learning, 3076–3085.

Simester D, Timoshenko A, Zoumpoulis SI (2020) Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. Management Science 66(6):2495–2522.

Stripling E, vanden Broucke S, Antonio K, Baesens B, Snoeck M (2018) Profit maximizing logistic model for customer churn prediction using genetic algorithms. Swarm and Evolutionary Computation 40:116–130.

Verbeke W, Baesens B, Bravo C (2017) Profit driven business analytics: a practitioner's guide to transforming big data into added value (John Wiley & Sons).

Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B (2012) New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218(1):211–229.

Verbraken T, Bravo C, Weber R, Baesens B (2014) Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research 238(2):505–513.

Verbraken T, Verbeke W, Baesens B (2012) A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Transactions on Knowledge and Data Engineering 25(5):961–973.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113(523):1228–1242.