Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 225 (2023) 2163–2172

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

Interpretable machine learning for heterogeneous treatment effect estimators with Double ML: a case of access to credit for SMEs

Kyrylo Medianovskyi[a,*], Aidas Malakauskas[b], Ausrine Lakstutiene[b], Sadok Ben Yahia[a]

[a] *Tallinn University of Technology, Ehitajate Street 5, Tallinn 19086, Estonia*
[b] *Kaunas University of Technology, K. Donelaičio Street 73, Kaunas 44029, Lithuania*

Abstract

Asymptotically consistent estimators of a treatment effect under many potential confounders became possible with the latest advancements in doubly-robust causal inference models (e.g., Double ML). In this study, we propose SAFE-TH framework to estimate and explain the heterogeneous treatment effect with partial dependence plots and report it under a reduced hypothesis space of interest. We analyze a shift in accessibility to credit for small to medium enterprises (SMEs) during the first months of the COVID-19 pandemic. Utilizing the proposed framework can improve the interpretability of CATE models by identifying and providing confidence intervals for regions of heterogeneity.

Keywords: Interpretable Machine Learning; Explainable Artificial Intelligence; Double ML; CATE; SHAP; Partial Dependence Plot

1. Introduction

Modern machine learning models are getting commonly adopted in policy making and decision support in econometrics [3], social sciences [22], and healthcare applications [32]. The problem of policy optimization can rarely be reduced to a prediction task. It requires counterfactual reasoning, e.g., “if we change X, the outcome Y will become n units higher.” Thus, the problem includes a causal aspect.

Estimating a treatment effect or uplift in business applications is not trivial. Each entity can be either treated or not, and the actual individual effect is never observed. Controlled settings of a randomized experiment with a defined treatment policy are not typical for real-world observational data. For the latter case, one would need to estimate a set of corrections that are not necessarily of interest to reduce the bias of a predictive machine learning model. For example, the propensity score tries to predict the likelihood of treatment based on observational data.

[*] Corresponding author.
E-mail address: kyrylo.medianovskyi@taltech.ee

10.1016/j.procs.2023.10.207

Recent advancements in interpretable machine learning (IML), interchangeably referenced as explainable artificial intelligence (XAI), [2] help to gain insights into the purely predictive black-box estimators (e.g., gradient-boosted trees, random forest, artificial neural networks). Such highly flexible models can provide a better fit for a potentially complex heterogeneity of treatment effects. At the same time, IML provides tools to investigate models from a specific point of interest: on the global (e.g., feature-wise) or data instance level. A rigorous form of causal inference is not available for black-box models, but IML can inform a procedure for feature reduction towards a smaller space of interest. Such a reduction opens an opportunity for linear treatment effect models that allow the construction of confidence intervals. As well, the advances of XAI can address the challenges faced by Small and Medium-sized Enterprises (SMEs) by providing insights into factors that determine their ability to access credit and by facilitating the evaluation of large quantities of non-linear factors through state-of-the-art machine learning algorithms.

SMEs are crucial contributors to the global economy's employment, innovation, and economic development [26]. Despite their importance, SMEs face daily challenges related to accessing credit and financial support, impeding their ability to export [25], innovate [8], and remain competitive [14]. These challenges may include lender-specific issues, such as the inability to properly evaluate opaque SMEs, macro-specific factors like market conditions, and individual company-specific factors, such as limited collateral availability, business volatility, and financial literacy [6]. Although the effect of individual factors has been studied using traditional modeling techniques, the evaluation of large quantities of factors and the non-linear nature of the problem make the task complicated and require state-of-the-art machine learning algorithms [12].

1.1. Related work

Estimators considering heterogeneity usually relate to conditional average treatment effect (CATE) models. CATE inputs a set of confounders (features) to estimate the outcome. A new sub-domain emerged in causal research proposing leveraging non-parametric or highly parameterized (deep learning) models in constructing CATE estimation. We refer to [9] and [21] for a review of meta-learners – algorithms that do not restrict the choice of an ML model to estimate CATE; and to a chain of work on the Double ML framework [11, 33, 16] that provides an unbiased treatment effect estimation allowing a relaxed model choice. As well, there is extensive work on model-specific CATE estimators: Gaussian processes [1], tree ensembles [4, 19], deep learning [20, 34, 38]. Interpretability of black-box CATE models is present in econometric research [37], in healthcare applications [10, 27]. It was proposed by [5] to perform rule mining using tree ensembles to obtain an interpretable CATE model. As well, Policytree [35] method proposes to provide interpretation of estimated treatment effects by fitting a small decision tree.

In contrast to existing CATE interpretation methods, our approach includes a supportive stage for experts to identify the treatment heterogeneity hypotheses of interest. For the decision support we leverage the performance of boosted, but opaque, estimators by providing a global interpretation with confidence intervals (CI) using bootstrap method. Bootstrap estimation of CI for a decision tree splitting parameters would be an unstable task due to the algorithm's sensitivity to changes in data subsets and general non-parametric design. We believe feature-wise global interpretation methods (e.g., partial dependence plots) with CI, that applied to boosted models, can provide a wider view of the backbone data generation process than the shallow models.

1.2. Contribution and paper's outline

We propose a framework to search for a reduced hypothesis space of treatment heterogeneity using model-agnostic interpretability methods – Supervised Assisted Feature Extraction for Treatment Heterogeneity (SAFE-TH). The partial dependence (PD) plots estimated with CI from the learning model perspective provide substance to support an expert's reasoning behind a simplifying parametric feature transformation. The reduced heterogeneity hypotheses and a measure of uncertainty provided by this method can be used to create interpretable treatment policies.

The remaining paper is organized as follows: Section 2 includes a description of the proposed framework and supplementary procedures, as well as details of the SME's dataset are presented; in section 3, we report the experiment setup, validity, and results; section 4 discusses findings; and in section 5 conclusion is drawn.

2. Method

This section describes the dataset used in this study and the proposed framework's methods: Double ML, partial dependence plots, and feature importance.

2.1. Data

For the study, we obtained a proprietary dataset of Lithuanian SMEs that applied for a loan between March 2019 and May 2020. We considered the Covid-19 shock to be introduced after 14 March 2020. $1,807$ applications collected after that date were regarded as "treated," and $5,188$ applications before were assigned to a control group. We took a target variable to indicate whether the application was rationed (rejected). Around 49% of the instances in the dataset had a positive label, meaning those were rejected. Table 1 shows the twenty-eight features that were added to the dataset, including six binary and five categorical.

Table 1. Description of Lithuanian SME dataset. Features with * are included for both current and previous years.

Feature name	Description
return_on_assets*	operating_profit / total_assets
current_gross_margin_ratio	(net_sales - cost_of_goods_sold) / net_sales
current_current_ratio	current_assets current_liabilities
current_debt_ratio	total_liabilities / total_assets
Ext_ovd_amt_2years	log_{10} of external overdue amount in EUR
Int_Ovd_Amt_last2years	log_{10} of internal overdue amount in EUR
SH_Ext_ovd_amt_2years	log_{10} of shareholders' external overdue amount in EUR
SH_Int_Ovd_Amt_last2years	log_{10} of shareholders' internal overdue amount in EUR
rejections	(binary) 1 if there was a rejection previously
diversity	(binary) 1 if the fraction of male/female minority of shareholders greater than 0.1
had_default_shareholder	(binary) 1 if one of the shareholders had a default
had_default	(binary) 1 if a company had a default
E-commerce	(binary) 1 if a company has had e-commerce product
POS	(binary) 1 if a company has used point of sales product
current_asset_turnover_ratio	net_sales / total_assets
current_receivables_turnover_ratio	net_sales / accounts_receivable
urban_rural	(categorical) 0-biggest city, 1-other towns, 2-rural area
current_tangible_assets_ratio	tangible_assets / total_assets
share_of_payment_transactions	incoming_cash_flow / net_sales
change_in_sales	(current_net_sales - previous_net_sales) / previous_net_sales
Nbr_fin_contracts	how many credit contracts a company already had
debt_share	counter_party_balance_amt / total_liabilities
relationship_duration	length of a relationship duration with a bank in days
segmentation	size of an enterprise (categorical) 0-micro, 1-small, 2-medium
Sector_group	(categorical) 0-Agriculture & Forestry & Fishing, 1-Commerce, 2-Construction, 3-Hotels & Restaurants, 4-Information & Communication, 5-Manufacturing, 6-Professional Services, 7-Real Estate, 8-Transportation
Legal_form	(categorical) 0-individual enterprise, 1-partnership, 2-private limited liability company
Product	(categorical) 0-credit card, 1-investment financing, 2-leasing, 3-trade finance, 4-working capital financing

2.2. Double ML

The Double ML framework, as described by [11], is a method used to estimate causal effects in machine learning. It involves modeling the relationship between confounding variables, treatment assignment, and outcome, and then using optimization techniques to estimate the conditional average treatment effect. By fitting models for confounding effects and treatment assignment, and combining them with weighted regression, the framework provides estimates of individual treatment effects that can be used for causal inference. Double ML framework from [11] can be described in terms of a data generation process (1) following [31]. It consists of several components: $G(X, W)$ – confounding or prognostic effect; $F(X, W)$ is a propensity score and T is either 0 or 1 in a binary treatment case; $\theta(X)$ – conditional average treatment effect (CATE). X represents confounders – features that influence both outcome and treatment effect. W are controls - the treatment effect does not depend on those features or is not of interest.

$$\begin{aligned} Y &= \theta(X) \cdot T + G(X, W) + \epsilon \,, \mathbb{E}[\epsilon \mid X, W] = 0 \\ T &= F(X, W) + \eta \,, \mathbb{E}[\eta \mid X, W] = 0, \quad \mathbb{E}[\eta \cdot \epsilon \mid X, W] = 0 \end{aligned} \tag{1}$$

In a set of observed values, $\{(X, W, T, Y)\}_n^{(i)}$, treatment T is either prescribed or not, which leads to incomplete information about its true effect for each case i. It was shown by [11] that under the unconfoundedness assumption for a set of observed values (X, W, T, Y) the CATE θ can be estimated in a doubly-robust setting and be asymptotically consistent in the case of being a constant or a low-dimensional linear function of X. The resulting optimization problem (2) requires both fitted prognostic model $\hat{G}$ and propensity score model $\hat{F}$.

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}\left[\left((Y - \hat{G}(X, W)) - (T - \hat{F}(X, W)) \cdot \theta(X)\right)^2\right] \tag{2}$$

It was proposed by [16] to use arbitrary models to fit $\hat{\theta}$. For a non-parametric estimation and in the case of a binary treatment, the optimization problem can be rewritten as (3), which can be solved with a weighted regression.

$$\hat{\theta} = \operatorname{argmin}_{\theta} \mathbb{E}\left[(T - \hat{F}(X, W))^2 \left(\frac{(Y - \hat{G}(X, W))}{(T - \hat{F}(X, W)} - \theta(X)\right)\right] \tag{3}$$

The resulting individual treatment effect (ITE) estimates are collected in a cross-validated manner. For each fold, $\hat{G}$ and $\hat{F}$ are fitted on the training subset, while the target ITE is calculated for a hold-out subset using those models. Finally, estimates of ITE from each fold are collected to fit the CATE model $\hat{\theta}$. According to [11], if θ is a low-dimensional parametric estimator, asymptotic confidence intervals for the parameters are allowed.

2.3. Partial dependence plots

Partial dependence (PD) plot [17] for a given estimator $\hat{f}$ and a feature of interest x_i is expressed as in (4), where x_{-i} stands for all other features except x_i. With a prediction set of input vectors $\mathbf{x}^{(j)}$ of size n, the PD-plot is estimated with (5).

$$PD(x_i) = \mathbb{E}_{x_{-i}}[\hat{f}(x_i, x_{-i})] \tag{4}$$

$$\widehat{PD}(x_i) = \frac{1}{n} \sum_{j=1}^{n} \hat{f}(x_i, x_{-i}^{(j)}) \tag{5}$$

Such a method estimates the dependence between the feature of interest and the outcome provided by a model $\hat{f}$. Because the estimation of PD-plots relies on a prediction set, it can be unreliable in a less represented region of x_i, e.g., extreme edge values of x_i. A learner-PD [28] method proposes point-wise confidence intervals for PD-plots. To estimate the variance (6), several models $\hat{f}_d$ are fitted on bootstrapped sub-samples of data. $\overline{\widehat{PD}}$ is an average of $\widehat{PD}_d$ estimators over m refits.

$$\hat{\mathbb{V}}(\overline{\widehat{PD}}(x)) = \left(\frac{1}{m} + c\right) \cdot \frac{1}{m-1} \sum_{d=1}^{m} \left(\widehat{PD}_d(x) - \overline{\widehat{PD}}(x)\right)^2 \tag{6}$$

A correction term $c = \frac{n_2}{n_1}$ is proposed by [29], where n_1 and n_2 correspond to sizes of train and test sets. This term attempts to compensate for the fact that bootstrap refits share data across sub-samples and lead to a true variance underestimation in a naive setting ($c = 0$). The learner-PD confidence intervals are estimated as in (7), where $t_{1-\frac{\alpha}{2}}$ corresponds to a $1 - \frac{\alpha}{2}$ quantile of t-distribution with $m - 1$ degrees of freedom.

$$CI_{\overline{\widehat{PD}}(x)} = \left[\overline{\widehat{PD}}(x) - t_{1-\frac{\alpha}{2}} \sqrt{\hat{\mathbb{V}}(\overline{\widehat{PD}}(x))};\ \ \overline{\widehat{PD}}(x) + t_{1-\frac{\alpha}{2}} \sqrt{\hat{\mathbb{V}}(\overline{\widehat{PD}}(x))}\right] \tag{7}$$

2.4. Feature importance

Feature importance methods try to attribute each input variable according to some score. One of the methods is Permutation Feature Importance (PFI) [7, 15]. PFI tests a feature's ability to improve the model performance regarding a chosen metric, e.g., mean square error or cross-entropy. The score of feature X_i is calculated as the difference between a metric estimated over an original dataset and a modified dataset with randomly perturbed X_i. Another important attribution method is SHapley Additve exPlanations (SHAP) [23]. SHAP works on a single instance of a dataset. It uses a game-theory approach to ensure that all of the features in the input vector have an equal impact on the model's outcome. For each feature in the prediction set, the mean absolute SHAP value can be used to figure out the global-level importance. Like PD-plot, SHAP quantifies the effect of a feature on the model's outcome. That contrasts with PFI, which indicates a contribution to performance and requires true labeling of the outcome. Both methods cannot be regarded as reliable from a causal perspective if the model is under-fitted or the features are correlated.

2.5. Proposed framework

Similar to SAFE [18] (Supervised Assisted Feature Extraction), we propose to use PD to simplify continuous features into a piece-wise constant space. The idea of hypothesis space reduction for the CATE model was discussed in [36, 5]. If the new space is parametric, it leads to asymptotically normal parameter estimates and, thus, facilitates the construction of confidence intervals. Such a method also helps to capture the heterogeneity of the treatment in an interpretable manner. We propose an extension to SAFE: SAFE for treatment heterogeneity (SAFE-TH) that includes four steps, as shown in Algorithm 1. The proposed method introduces extra control over the feature transformation as it is supplied with variance estimation for PD-plots.

Algorithm 1 Procedure for Supervised Assisted Feature Extraction for Treatment Heterogeneity (SAFE-TH).

```
1) Fit a treatment heterogeneity effect estimator.
2) Select the essential features of the treatment effect model and build PD-plots.
3) Choose a feature of interest from a set of the most important ones and transform it into one-hot binned space
   considering the information from the PD-plot.
4) Fit a new linear treatment heterogeneity effect model that only depends on a transformed interest feature and
   reports each estimated parameter's significance level and confidence interval.
```

3. Experiment and results

This section reports the experiment setup and results for each framework component. We tested the performance of confidence intervals for Learner-PD and feature importance ranking based on the synthetic and semi-synthetic data correspondingly. The final results include a description of the obtained piece-wise constant functions with asymptotic and bootstrap confidence interval estimates on the real dataset.

3.1. Model selection

We tested the performance of several ML algorithms over our dataset, including Lasso, Random Forest, and LightGBM (Table 2). The characteristics were tested and found to be unrelated. The absolute value of Spearman's correlation for ordered and continuous features and Matthew's coefficient for categorical features remained below 0.4, which corresponds to a below moderate correlation level. The treatment effect in the Double ML setting is expected to be linear regarding the prognostic effect and propensity score models. In the former case, we chose regressor versions of the tested methods to work around this limitation instead of classifiers to eliminate a logistic non-linear link function. We tuned greedily by fitting each of the models with a 5-fold cross-validation. We collected the mean value of average precision and the area under the receiver operating characteristic curve (ROC AUC) to compare models. To estimate both measures, we cut the output score to fit between 0 and 1. The number of estimators was limited to 250 for the

Random Forest model and 300 for the LightGBM model. Furthermore, to prevent over-fitting and complex interactions, we limited the maximum tree depth for the latter models to three, and the maximum number of leaves to eight. The learning rate of LightGBM was tuned down to 0.095. We took the LightGBM model as the best base learner for further experiments to fit our data well.

Table 2. Selection across possible models. Average CV score.

Model	Average Precision	ROC AUC
Lasso	0.570499	0.582207
Random Forest	0.763213	0.761838
LightGBM	0.796917	0.803652

3.2. Confidence intervals for learner-PD

We conducted a simulation study to empirically check the learner-PD method's coverage of the confidence interval (CI). With a data generation process (DGP) in (8) that mimics some of the qualities of our real dataset, 1, 000 synthetic samples of size 7, 000 were generated. We estimated the CATE model with a 3-fold cross-fit using a Double ML framework. According to a cross-validation setting, the CATE model is fitted to samples that combine the results of all folds. For CATE PD variance estimation, we used a conservative correction term of $c = 2/1$ for a case of 3-fold cross-validation, as the testing data for models $\hat{G}$ (prognostic) and $\hat{F}$ (propensity score) corresponds to the training data for $\hat{\theta}$ (CATE). We compared the average coverage and width of 90% CI depending on the number of bootstrap model refits (2 to 25); see Figure 1. For each important feature, we measured how frequently the ground-truth PD was within the learner-PD CI at five equidistant points within the box-plot whiskers for that feature. While the CI estimated with the correction term c reached the target 90% coverage, the CI without correction remained close to the 82% threshold.

$$\begin{aligned}
X_1 &= Poiss(0.5)\\
X_2 &= \mathcal{N}(0.5, 1)\\
X_3 &= Bernoulli(0.3)\\
X_4 &= Discrete(P(0) = 0.35, P(1) = 0.25, P(2) = 0.22, P(3) = 0.18)\\
X_5 &= U(0, 1)\\
X_{6-10} &= \mathcal{N}(0, 1)\\
X_{11-15} &= U(0, 1)\\
Y &= \theta(X) \cdot T + G(X) + \epsilon_y, \quad \epsilon_y \sim \mathcal{N}(0, 0.1^2)\\
T &= Bernoulli(0.2)\\
G(X) &= 0.5 - 0.02X_1 + 0.001X_2 + 0.1X_3 - 0.02X_5 \cdot \mathbf{1}\{X_5 > 0.2\}\\
\theta(X) &= 0.08 \cdot \mathbf{1}\{X_1 = 0\}\\
&+ (-0.05 \cdot \mathbf{1}\{X_2 \leq 0.3\} + 0.1 \cdot \mathbf{1}\{X_2 > 0.55\} + 0.03 \cdot \mathbf{1}\{X_2 > 1\})\\
&+ 0.15 \cdot \mathbf{1}\{X_4 = 3\}\\
&+ (-0.15X_6 + \epsilon_6 \cdot \mathbf{1}\{X_6 > 1\})\\
&\epsilon_6 \sim \mathcal{N}(0, 1)
\end{aligned} \tag{8}$$

3.3. Validation of feature importance

To validate the feature importance methods, we used a semi-synthetic dataset with the original features but artificially generated outcome and treatment effect, similar to [13], see (9). With the synthetic labels and treatment effects, we aimed to mimic the features' marginal behavior realistically. The Top-5 most important covariates of the treatment effect model were collected over 100 simulations. To rank them by the importance level, we applied the mean absolute SHAP, PFI, and split importance (calculated as the number of times a feature was used in a decision split of a tree-based model). For the mean absolute SHAP score, we used a TreeSHAP [24] implementation that utilizes an

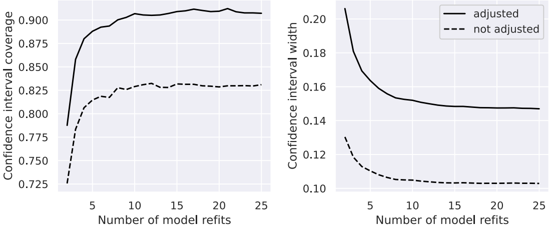

Fig. 1. Average learner-PD confidence interval coverage (left) and width (right) depending on the number of bootstraps refits for the CATE model for the fully synthetic DGP data.

inherent tree structure of a model to impute a background set. Also, for the PFI loss function, we took a negative mean squared error weighted according to the propensity score of each sample. In Table 3, top-5 coverage corresponds to the number of actually important features within the reported top-5; top correct reflects the number of correctly reported consecutive features before the first appearance of a wrong one. Both mean absolute SHAP and weighted PFI returned a median of 3 top-correct features, but SHAP demonstrated slightly better performance in terms of mean top-5 and top-correct values. The split importance provided the worst results.

$$\begin{aligned}
\theta_1(X) &= \mathbf{1}\{X_1 > 0.2\} \cdot 0.05 \cdot X_1 + \mathbf{1}\{X_1 > 0.8\} \cdot \epsilon_{\theta_1}, \quad \epsilon_{\theta_1} \sim \mathcal{N}(0, 0.2^2)\\
\theta_2(X) &= \mathbf{1}\{X_2 = 3\} \cdot (-0.05) + \mathbf{1}\{X_2 = 4\} \cdot 0.1\\
\theta_3(X) &= \mathbf{1}\{X_3 \leq 0.2\} \cdot (X_3 - 0.2) + \mathbf{1}\{X_3 > 0.65\} \cdot (X_3 - 0.65) \cdot 0.03\\
\theta_4(X) &= \mathbf{1}\{X_4 = 1\} \cdot (-0.02) + \mathbf{1}\{X_4 \in [1, 6]\} \cdot 0.1 + \mathbf{1}\{X_4 > 6\}\\
\theta_5(X) &= \mathbf{1}\{X_5 \in [0.7, 0.9]\} \cdot 0.1 + \mathbf{1}\{X_5 > 0.9\} \cdot (-0.1 + \epsilon_{\theta_5}), \quad \epsilon_{\theta_5} \sim \mathcal{N}(0, 0.1^2)\\
\theta_{synthetic} &= \textstyle\sum_{i=1}^{5} \theta_i(X)\\
G_1(X) &= 0.25 - 0.05X_4\\
G_2(X) &= \mathbf{1}\{X_2 = 0\} \cdot (-0.1) + \mathbf{1}\{X_2 = 1\} \cdot 0.3 + \mathbf{1}\{X_2 = 3\} \cdot 0.1 + \mathbf{1}\{X_2 = 4\} \cdot 0.1\\
G_3(X) &= \mathbf{1}\{X_5 > 0.02\} \cdot 0.05\\
G_4(X) &= \mathbf{1}\{X_6 \leq 0.7\} \cdot 0.01 \cdot X_6 + \mathbf{1}\{X_6 > 0.7\} \cdot 0.04 \cdot (X_6 - 0.7)\\
G_5(X) &= 0.15\, X_7\\
G_6(X) &= \mathbf{1}\{X_8 \leq 3650\} \cdot (0.05 - X_8 \cdot 0.05/3650)\\
G_{synthetic} &= \textstyle\sum_{i=1}^{6} G_i(X) + 0.22\\
Y_{synthetic} &= \mathbf{1}\{\theta_{synthetic} \cdot T + G_{synthetic} + \epsilon_y > 0.4\}, \quad \epsilon_y \sim \mathcal{N}(0, 0.1^2)
\end{aligned} \tag{9}$$

In (9) X_1 corresponds to ‘current_tangible_assets_ratio’ , X_2 to ‘Product‘ , X_3 to ‘current_gross_margin_ratio’ , X_4 to ‘Nbr_fin_contracts’ , X_5 to ‘debt_share’ , X_6 to ‘current_debt_ratio’ , X_7 to ‘rejections’ , X_8 to ‘relationship_duration’. Exposure to Covid-19 shock T remained identical to the real dataset. $Y_{synthetic}$ was intentionally transformed into a binary variable to imitate the original label.

Table 3. Comparison of feature importance methods over semi-synthetic data. Average over 100 simulations.

Importance	mean top-5 coverage	mean top correct	median top correct
Mean Abs SHAP	3.55	3.30	3
PFI	3.29	2.92	3
Split	2.53	1.80	2

3.4. Obtained results from real-life datasets

We fitted a Double ML framework with a LighGBM regressor for the prognostic model, the same type of regressor for the CATE model, and the LightGBM classifier for the propensity score model. A 3-fold cross-fit was

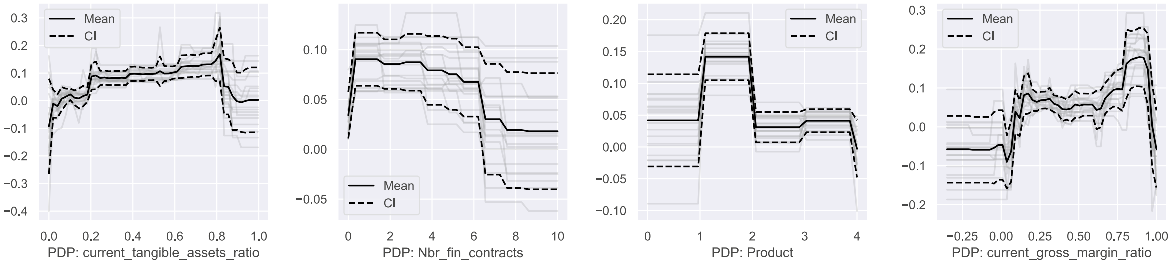

Fig. 2. Learner-PD plots for the top four most important features with 90% CI; y-axis represents the estimated average model prediction conditioning on the value of the feature of interest; grey lines show PD for each refit.

used. According to the mean absolute SHAP of the CATE model, the first four most important features were 'current_tangible_assets_ratio', 'Nbr_fin_contracts', 'Product', 'current_gross_margin_ratio'. The learner-PD plots of over 20 refits with correction in Figure 2 provided insights to perform feature binning. Table 4 reports the significance of the piece-wise constant function for each feature as a single confounder of the CATE model. In the latter setting, prognostic and propensity score models remained the same, but the CATE model was fitted with weighted least squares corresponding to the binned dummy variables without intercept. We collected estimates of each naive and bootstrapped constant with 20 refits and correction ($c = 2$). The areas with lower variance in the PD plots inherently provided more significant constant estimates. For the naive method, we used a double-sided Z-test with a null hypothesis that the estimated value differs from zero. In case of the bootstrap estimation, we used the same hypothesis for a t-test. A p-value of the Jarque–Bera test was also reported to support the normality assumption of the values' distribution across bootstrapped refits.

Table 4. Findings for CATE heterogeneity over reduced hypothesis space. Top-4 important features with (0.05, 0.95) confidence interval quantiles. $P >$ stands for a p-value of a corresponding test.

		Naive			Bootstrap			
Feature	Bin	Value	$P > \mid Z \mid$	Asymptotic CI	Value	$P > \mid t \mid$	$P > JB$	Bootstrap CI
current_tangible_assets_ratio	≤0.2	0.010	0.583	[-0.020, 0.041]	0.008	0.055	0.700	[-0.035, 0.050]
	(0.2, 0.9]	0.110	< 0.001	[0.083, 0.138]	0.108	<0.001	0.754	[0.080, 0.136]
	> 0.9	-0.129	0.268	[-0.319, 0.062]	-0.105	<0.001	0.366	[-0.385, 0.175]
Nbr_fin_contracts	≤0.5	0.015	0.455	[-0.018, 0.048]	0.013	0.003	0.969	[-0.030, 0.056]
	(0.5, 6.5]	0.100	< 0.001	[0.070, 0.129]	0.099	<0.001	0.814	[0.058, 0.140]
	> 6.5	0.011	0.723	[-0.040, 0.063]	0.001	0.851	0.862	[-0.071, 0.074]
Product	{0,1,2}	0.037	0.013	[0.012, 0.061]	0.036	<0.001	0.961	[0.000, 0.072]
	{3}	-0.069	0.121	[-0.142, 0.004]	-0.069	<0.001	0.858	[-0.163, 0.025]
	{4}	0.173	< 0.001	[0.130, 0.217]	0.174	<0.001	0.766	[0.093, 0.254]
current_gross_margin_ratio	≤ 0.0	-0.057	0.473	[-0.189, 0.074]	-0.018	0.223	0.845	[-0.178, 0.141]
	(0.0, 0.8]	0.062	< 0.001	[0.041, 0.084]	0.058	<0.001	0.889	[0.031, 0.085]
	>0.8	0.091	0.073	[0.007, 0.174]	0.107	<0.001	0.487	[0.009, 0.205]

4. Discussion

In this study, we introduce a novel framework aimed at exploring a narrowed-down hypothesis space of treatment heterogeneity by leveraging model-agnostic interpretability techniques. Specifically, we employ partial dependence (PD) plots, accompanied by confidence intervals derived from the learning model's perspective, to substantiate an expert's rationale behind employing a simplified parametric feature transformation. The utilization of this approach offers two key advantages. Firstly, it facilitates the identification of reduced heterogeneity hypotheses, thereby enabling a more focused analysis of treatment effects. Secondly, it provides a measure of uncertainty that enhances the interpretability of the obtained results.

Based on the modeling results, as demonstrated in Table 4 and Figure 2, it is evident that the importance and confidence intervals for individual features are not equally distributed across all feature values. The uncertainty in PD-plots corresponds to the uncertainty in the piece-wise constant hypothesis space because the PD-plot shows how the average predicted value of the target variable changes as the input variables are varied. The piece-wise constant hypothesis space determines the regions of the input space where the model assumes a constant value for the target variable. Therefore, the uncertainty in the PD-plot reflects the uncertainty in the model's decision rules and the regions of the input space where the model's assumptions hold. Consequently, these reduced heterogeneity hypotheses and the associated uncertainty measure can be effectively utilized to formulate treatment policies that are comprehensible and explainable to domain experts.

By applying the framework for the case of access to credit for SMEs and creating individual bins for highlighted features, it is evident that practical implications exist. All displayed confounders contain value regions of higher and lower uncertainty. Specifically, it is exacerbated at extremes. Throughout most values (bin [0.2–0.9]), the current tangible asset ratio demonstrates a relatively low impact on the model's prediction. A slight upward trend indicates that companies with higher values were more likely to be credited rationed during times of uncertainty. At high feature values (bin [>0.9]), the effect of a higher current tangible asset ratio starts to impact a company's ability to access credit positively. Though the average effect is positive, the result is inconclusive, as the CI are wide and indicate both positive and negative impacts on overall credit accessibility during times of uncertainty. Findings for the current gross margin ratio (similar to the current tangible assets ratio) suggest that the model is relatively well able to estimate the impact on the model's outcome for center values (bin [0.0–0.8]) while decreasing in certainty for tail values. These findings suggest that even if the contribution of a feature is important, the extent might not be uniform across actual feature values. Based on the provided insights, a practical use of the proposed framework could be applied in real world decision making, where feature values with narrow CI would provide conclusive insights, while wider CI would indicate higher uncertainty and require further evaluation.

5. Conclusion

The speed and efficiency of modern supervised ML models under the hood of Double ML provide an opportunity to estimate a treatment effect with many control and confounding features. The proposed SAFE-TH framework helps to leverage the interpretability methods for heterogeneous treatment effect estimators and allocate a smaller piecewise constant hypothesis space for the features of interest. The reported significance of each binning threshold over the given confounder helps to provide interpretable treatment policies.

The provided estimation of bootstrapped PD point-wise CI does not inform a spread in the data. While the percentile CI method could help, that would require at least 100 model recalculations over bootstrapped samples. Despite that, the resulting bootstrap CI of heterogeneity coefficients is trustworthy due to their asymptotic normality. The first three months of the pandemic were chosen as a treatment period for SMEs. This study was limited in data to observe further possible effects through time, including the consecutive supportive measures.

Future work might extend the treatment effect estimation with non-linearity for the outcome model proposed by [30], specifically to use a logistic link function in the case of a binary dependent variable. As well, the framework might include methods to search for feature interactions.

References

[1] Alaa, A.M., Van Der Schaar, M., 2017. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. Advances in Neural Information Processing Systems 2017-December, 3425–3433. arXiv:1704.02801.

[2] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion doi:10.1016/j.inffus.2019.12.012.

[3] Athey, S., Imbens, G.W., 2017. The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives 31, 3–32. doi:10.1257/jep.31.2.3.

[4] Athey, S., Tibshirani, J., Wager, S., 2016. Generalized Random Forests. Annals of Statistics 47, 1179–1203. doi:10.1214/18-AOS1709.

[5] Bargagli-Stoffi, F.J., Cadei, R., Lee, K., Dominici, F., 2020. Causal Rule Ensemble: Interpretable Discovery and Inference of Heterogeneous Treatment Effects arXiv:2009.09036.

[6] Berger, A.N., Udell, G.F., 2006. A more complete conceptual framework for sme finance. JOURNAL OF BANKING & FINANCE 30, 2945–2966. doi:10.1016/j.jbankfin.2006.05.008.

[7] Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

[8] Brown, J.R., Martinsson, G., Petersen, B.C., 2013. Law, stock markets, and innovation. The Journal of Finance 68, 1517–1549. doi:10.1111/jofi.12040.

[9] Caron, A., Baio, G., Manolopoulou, I., 2022a. Estimating individual treatment effects using non-parametric regression models: A review. Journal of the Royal Statistical Society. Series A: Statistics in Society 185, 1115–1149. doi:10.1111/rssa.12824.

[10] Caron, A., Baio, G., Manolopoulou, I., 2022b. Interpretable Deep Causal Learning for Moderation Effects arXiv:2206.10261.

[11] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21, C1–C68. doi:10.1111/ectj.12097.

[12] Correa Bahnsen, A., Aouada, D., Stojanovic, A., Ottersten, B., 2016. Feature engineering strategies for credit card fraud detection. Expert Systems with Applications 51, 134–142. doi:10.1016/j.eswa.2015.12.030.

[13] Crabbé, J., Curth, A., Bica, I., van der Schaar, M., 2022. Benchmarking Heterogeneous Treatment Effect Models through the Lens of Interpretability , 9–13arXiv:2206.08363.

[14] Ferrando, A., Ruggieri, A., 2018. Financial constraints and productivity: Evidence from euro area companies. International Journal of Finance & Economics 23, 257–282. doi:10.1002/ijfe.1615.

[15] Fisher, A., Rudin, C., Dominici, F., 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research 20. arXiv:1801.01489.

[16] Foster, D.J., Syrgkanis, V., 2019. Orthogonal Statistical Learning arXiv:1901.09036.

[17] Friedman, J.H., 1991. Multivariate Adaptive Regression Splines. The Annals of Statistics 19, 1–23. doi:10.1214/aos/1176347963.

[18] Gosiewska, A., Kozak, A., Biecek, P., 2021. Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. Decision Support Systems 150, 113556. doi:10.1016/j.dss.2021.113556.

[19] Hahn, P.R., Murray, J.S., Carvalho, C.M., 2020. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). Bayesian Analysis 15, 965–1056. doi:10.1214/19-BA1195.

[20] Johansson, F.D., Shalit, U., Sontag, D., 2016. Learning representations for counterfactual inference. 33rd International Conference on Machine Learning, ICML 2016 6, 4407–4418. arXiv:1605.03661.

[21] Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B., 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences of the United States of America 116, 4156–4165. doi:10.1073/pnas.1804597116.

[22] Leist, A.K., Klee, M., Kim, J.H., Rehkopf, D.H., Bordas, S.P., Muniz-Terrera, G., Wade, S., 2022. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. Science Advances 8. doi:10.1126/sciadv.abk1942.

[23] Lundberg, S., Lee, S.I., 2017. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems arXiv:1705.07874.

[24] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2, 56–67. doi:10.1038/s42256-019-0138-9.

[25] Manova, K., Wei, S.J., Zhang, Z., 2015. Firm exports and multinational activity under credit constraints. The Review of Economics and Statistics 97, 574–588. doi:10.1162/REST_a_00435.

[26] Manzoor, F., Wei, L., Siraj, M., 2021. Small and medium-sized enterprises and economic growth in pakistan: An ardl bounds cointegration approach. Heliyon 7, e06340. doi:10.1016/j.heliyon.2021.e06340.

[27] Meid, A.D., Gerharz, A., Groll, A., 2022. Machine learning for tumor growth inhibition: Interpretable predictive models for transparency and reproducibility. CPT: Pharmacometrics & Systems Pharmacology 11, 257–261. doi:10.1002/psp4.12761.

[28] Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M.N., Bischl, B., 2021. Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process arXiv:2109.01433.

[29] Nadeau, C., Bengio, Y., 2003. Inference for the Generalization Error. Machine Learning 52, 239–281. doi:10.1023/A:1024068626366.

[30] Nekipelov, D., Semenova, V., Syrgkanis, V., 2021. Regularised orthogonal machine learning for nonlinear semiparametric models. The Econometrics Journal 25, 233–255. doi:10.1093/ectj/utab022.

[31] Robinson, P.M., 1988. Root-n-consistent semiparametric regression. Econometrica: Journal of the Econometric Society , 931–954.

[32] Sanchez, P., Voisey, J.P., Xia, T., Watson, H.I., O'Neil, A.Q., Tsaftaris, S.A., 2022. Causal machine learning for healthcare and precision medicine. Royal Society Open Science 9. doi:10.1098/rsos.220638.

[33] Semenova, V., Goldman, M., Chernozhukov, V., Taddy, M., 2017. Estimation and Inference on Heterogeneous Treatment Effects in High-Dimensional Dynamic Panels under Weak Dependence arXiv:1712.09988.

[34] Shalit, U., Johansson, F.D., Sontag, D., 2017. Estimating individual treatment effect: Generalization bounds and algorithms. 34th International Conference on Machine Learning, ICML 2017 6, 4709–4718. arXiv:1606.03976.

[35] Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., Wager, S., 2020. policytree: Policy learning via doubly robust empirical welfare maximization over trees. Journal of Open Source Software 5, 2232. doi:10.21105/joss.02232.

[36] Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., Lewis, G., 2019. Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments. Advances in Neural Information Processing Systems 32. arXiv:1905.10176.

[37] Yang, J.C., Chuang, H.C., Kuan, C.M., 2020. Double machine learning with gradient boosting and its application to the Big N audit quality effect. Journal of Econometrics 216, 268–283. doi:10.1016/j.jeconom.2020.01.018.

[38] Yao, L., Huai, M., Li, S., Gao, J., Li, Y., Zhang, A., 2018. Representation learning for treatment effect estimation from observational data. Advances in Neural Information Processing Systems 2018-December, 2633–2643.