IMT School for Advanced Studies, Lucca
Lucca, Italy

KU Leuven, Faculty of Economics and Business
Leuven, Belgium

Essays on Applied Machine Learning

Joint degree in “PhD in Institutions, Markets and Technologies - Curriculum in Economics, Management and Data Science” and the Doctoral Program in “Economics”

XXXII Cycle

By

Falco J. Bargagli-Stoffi

2020

The dissertation of Falco J. Bargagli-Stoffi is approved.

PhD Program Coordinator: Prof. Dr. Pietro Pietrini, IMT School for Advanced Studies Lucca

Supervisors:
Prof. Dr. Kristof De Witte, KU Leuven, Maastricht University
Prof. Dr. Massimo Riccaboni, IMT School for Advanced Studies Lucca

The dissertation of Falco J. Bargagli-Stoffi has been reviewed by:

Prof. Dr. Giorgio Gnecco, IMT School for Advanced Studies Lucca
Prof. Dr. Fabrizia Mealli, University of Florence
Prof. Dr. Rachel Nethery, Harvard University
Prof. Dr. Mike Smet, KU Leuven

IMT School for Advanced Studies Lucca
KU Leuven, Faculty of Economics and Business
2020

To Alessandra, Giosiana and Pier Francesco

Contents

List of Figures

List of Tables

Acknowledgements

I consider myself a very fortunate person. In my life, I have met and I have been surrounded by wonderful people that have helped me grow as a PhD student, as a researcher, and, most importantly, as an individual. First, I would like to express my deepest gratitude to my Supervisors *Prof. Dr. Kristof De Witte* and *Prof. Dr. Massimo Riccaboni*. Thank you for your advises, your guidance and for providing me with all I have needed to navigate through my PhD. *Kristof*, thank you for your constant guidance, for your openness to explore new topics and research ideas, for the great opportunity of joining your team and for making me feel at ease in every moment. *Massimo*, thank you for taking me under your wing from the beginning of my journey as a PhD student, for all the opportunities that you offered me and for your plain-spoken advice, feedback and suggestions. Furthermore, I would not have enrolled in a PhD and nor would I have written this Dissertation without the mentorship of *Prof. Dr. Fabrizia Mealli*. *Fabrizia*, thank you for guiding and supporting me in making of one of the best choices of my life, for always finding the time to debate with me, and for your ceaseless guidance in the last six years. A great influence on my research has come from *Prof. Dr. Francesca Dominici*. *Francesca*, thank for your unceasing and enthusiastic support, for the opportunity of being involved in projects that helped me grow a lot as a researcher, and for making me feel at home in your (extended) team.

I want to thank the external referees *Prof. Dr. Rachel Nethery*, *Prof. Dr. Mike Smet* and the members of my committee *Prof. Dr. Giorgio Gnecco* and *Prof. Dr. Fabrizia Mealli*. The quality of my Dissertation has improved considerably owing to your criticisms, comments, and suggestions. My gratitude also to my co-authors, *Prof. Dr. Gustavo Cevolani, Kenneth De Beckker, Prof. Dr. Laura Forastiere, Prof. Dr. Kwonsang Lee, Joana Maldonado, Dr. Jan Niederreiter, Prof. Dr. Armando Rungi* and *Costanza Tortú*. I would also like to thank every researcher and professor who have contributed to my work and my personal growth as a researcher. In particular, *Prof. Dr. Giorgio Gnecco* for his constant supervision since I moved my very first steps in academia and to *Prof. Dr. Alberto Gherardini* for assisting me when I was a sprout BSc student looking for interesting research topics. I am grateful to *Chiara Magini, Maria Mateos* and *Sara Olson* in Lucca, *Ann Van Espen* in Leuven, and *Susan Luvisi* in Boston for their excellent administrative support during my PhD. None of my (many) visiting periods, would have been possible without the 24/7 help of the tireless *Daniela Giorgetti*.

When I started my PhD in Lucca, I was hoping to find nice and friendly colleagues, but I was not expecting to build the deepest bonds of friendship. To *Abhishek, Alessandra, Anna, Costanza, Federica, Francesco, Pietro, Nicolò, Sara, Tatiana*. To my friends and colleagues in Leuven, *Anna, Deni, Fritz, Francisco, Joana, Kaat, Kenneth, Mara, Silvia, Vita* and in Boston, *Ben, Danielle, Giovanni, Jenny, Rachel, Kwonsang, Xiao*. To *Giovanna*, for being the best mentor I could have. This journey would have been much harder without your support, your help, and the amazing moments we spent together.

I wrote this Dissertation in 5 different countries and across 2 continents. I can't be more grateful to share my path with incredible friends that have always found a way to support me even when I was far away. Thank you for constantly sending me your love from Champs sur le Bisance - *Francesco, Leonardo, Luigi, Gianni, Gianmarco, Giacomo, Matteo* - from Marcialla City - *Dario, Salvo, Marta, Matteo* - and from all around the world - *Akkie, Anna, Burak, Chris, Cosimo, Fabio, Flavia, Isabelle, Hans, Paul, Lucrezia, Madalina, Naomi, Vera and Vlad*. To *Lara*, for sharing with me most of this path.

Finally, I express my deepest gratitude to all my family, uncles and cousins for their support and affection throughout all my life. Foremost, I want to thank the three people this Dissertation is dedicated to. To my mother, *Alessandra*, for indicating me the way. To my father, *Pier Francesco*, for making the way wiser, lighter, and funnier. To my grandma, *Giosiana*, for unconditionally sustaining me in every step I took on the way.

Vita

November 14th, 1991	Born, Florence (Tuscany), Italy
Education	
2018–	Ph.D. candidate in Economics Faculty of Economics and Business, KU Leuven, Leuven, Belgium
2016–	Ph.D. candidate in Economics, Management and Data Science (XXXII cycle) IMT School for Advanced Studies Lucca, Lucca, Italy
2019	Visiting scholar Harvard University, Cambridge, Massachusetts, USA
2016	M.Sc. in Statistical, Financial and Actuarial Sciences Final mark: 110/110 summa cum laude University of Florence, Florence, Italy
2016	Visiting student Faculty of Economics and Business, Humboldt University, Berlin, Germany
2014	B.Sc. in Sociology and Social Policies Final mark: 110/110 summa cum laude University of Florence, Florence, Italy

Please find an up-to-date version of my Curriculum Vitae at the following address: `www.falcobargaglistoffi.com`

Academic Appointments

2018– Junior Researcher,
Leuven Economics of Education Research,
Faculty of Economics and Business, KU Leuven,
Leuven, Belgium

2016– Research Assistant,
Laboratory for the Analysis of Complex Economic Systems, IMT School for Advanced Studies Lucca,
Lucca, Italy

2019 Research Assistant,
Chan School of Public Health, Harvard University,
Cambridge, Massachusetts, USA

Teaching

IMT School (graduate courses)

2019– Lecturer for Data Science Lab

2018 Teaching Assistant for Machine Learning

Maastricht University (graduate course)

2020 Machine learning for social sciences

Flemish Ministry of Education

2020 Machine learning for education policies

Additional Education

2014 Advanced Training Program in Administrative Law, Scuola Superiore Sant'Anna, Pisa, Italy.

Honors, Grants and Awards

1. Ph.D. Full Scholarship in Economics, Faculty of Economics and Business, KU Leuven, Leuven, Belgium
2. INdAM-GNAMPA (Italian National Institute for High Mathematics, National Group for Mathematical Analysis, Probability, and their Applications) 2020 Project Funding
3. ERASMUS+ Mobility Consortium Grant (Academic Year 2019/2020)
4. Visiting Scientist Scholarship, Harvard University
5. Ph.D. Full Scholarship in Economics (XXXII cycle), IMT School for Advanced Studies Lucca, Lucca, Italy
6. ACIC 2019 Tom Ten Have Mention of Honour, Mc Gill University - University of Montreal
7. Galileo 2019 Project Funding
8. DSAA 2018 Student Travel Grant Award, IEEE Computational Intelligence Society
9. ERASMUS+ Mobility Consortium Grant (Academic Year 2018/2019)
10. Graduation Prize Bardazzi Mention of Honour, PIN - Polo Universitario di Prato
11. Graduation Prize Villa Favard, awarded by the Alumni association of the Faculty of Economics, University of Florence for the most original Master's thesis in the Academic Year 2015/2016
12. M.Sc degree with Honors
13. ERASMUS+ MOBILITY Grant (Academic Year 2015/2016)
14. B.Sc degree with Honors
15. Best Diploma Grade Award, Liceo Scientifico Agnoletti, Sesto Fiorentino, Florence, Italy
16. Mention in the Italian Excellence List, Albo Nazionale delle Eccellenze INDIRE

Publications

Peer reviewed publications

1. Bargagli-Stoffi, F. J., Gnecco, G. 2020 "Causal Tree with Instrumental Variable: an Extension of the Causal Tree Framework to Irregular Assignment Mechanisms." *International Journal of Data Science and Analytics*, 9, 315–337.
2. Bargagli Stoffi, F. J., Riccaboni M., Niederreiter, J. 2020 "Supervised Learning for the Prediction of Firm Dynamics", in S. Consoli, D. Reforgiato Recupero, M. Saisana (EDs) *Data Science for Economics and Finance: Methodologies and Applications*, Springer, forthcoming
3. Bargagli Stoffi, F. J., Cevolani, G., Gnecco, G. 2020 "Should Simplicity Be Always Preferred to Complexity in Supervised Machine Learning?" 6th International Conference on machine Learning, Optimization Data science (LOD). In *Lecture Notes in Computer Science*, Springer, forthcoming.
4. Bargagli-Stoffi, F. J., Gnecco, G. 2018 "Estimating Heterogeneous Causal Effects in the Presence of Irregular Assignment Mechanisms. *Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics* (DSAA).

Pre-prints and working papers

1. Bargagli-Stoffi, F. J., Tortù, C., Forastiere, L. 2020 "Heterogeneous Treatment and Spillover Effects under Clustered Network Interference.", working paper
2. Lee, K., Bargagli-Stoffi, F. J., Dominici, F. 2020 "Causal Rule Ensemble: Interpretable Inference on Heterogeneous Treatment Effects.", working paper
3. Bargagli-Stoffi, F. J., De Beckker, K., De Witte, K., Maldonado, J. 2020 "Assessing Sensitivity of Machine Learning Predictions. A Novel Toolbox with an Application to Financial Literacy.", working paper
4. Bargagli Stoffi, F. J., Riccaboni, M., Rungi, A. 2020 "Machine Learning for Zombie Hunting. Firms' Failures and Financial Constraints." *FEB Research Report Department of Economics* DPS20. 06.
5. Bargagli Stoffi, F. J., Cevolani, G., Gnecco, G. 2020 "Occam's Razor and Statistical Learning Theory", working paper
6. Bargagli-Stoffi, F. J., De Witte, K. Gnecco. G. 2019 "Heterogeneous Causal Effects with Imperfect Compliance: a Novel Bayesian Machine Learning Approach." *arXiv preprint* arXiv:1905.12707.

Presentations

1. St. Gallen University Invited Seminar, fall 2020, St. Gallen, Switzerland
2. Data Science for Impact Evaluation webinar (organizer), July 2020
3. International Conference on Machine Learning, Optimization, and Data Science, July 2020, Siena, Italy
4. Flemish Ministry of Education invited seminar February 2020, Brussels, Belgium
5. Bank of England/King's College conference on ”Big Data and Machine Learning: Interpretability and Model Uncertainty” November 2019, London, UK
6. Sant'Anna School for Advanced Studies Research Symposium, October 2019, Pisa, Italy
7. Harvard Data Science Initiative Research Symposium, May 2019, Cambridge, MA, USA
8. Atlantic Causal Inference Conference May 2019, Montréal, Canada
9. CEPR/EIEF/Bank of Italy conference on ”Firm Dynamics and Economic Growth” December 2018, Rome, Italy
10. Leuven Economics of Education Research Seminar December 2018, Leuven, Belgium
11. 5th International Conference on Data Science and Analytics October 2018, Turin, Italy
12. IMT School for Advanced Studies Research Symposium July 2018, Lucca, Italy
13. European Causal Inference Meeting April 2018, Firenze, Italy
14. Workshop on Recent Advances in Public Economics and Quantitative Methods March 2018, Lucca, Italy

Abstract

In this Dissertation, we deal with a series of applications of machine learning in the fields of social and health sciences. We introduce a set of novelties in the traditional usage of machine learning algorithms for predictive and causal inference tasks. In Part 1, we explore the field of machine learning for causal inference and we introduce two innovative techniques that combine state-of-the-art machine learning algorithms with causal inference methodologies. In the first Chapter, we introduce a novel Bayesian tree-based methodology to draw causal inference on heterogeneous effects in quasi-experimental scenarios. In the second Chapter, we account for possible drawbacks of tree-based methodologies by proposing a composite algorithm with a high level of interpretability and precision. In Part 2, we introduce applications of machine learning predictive power to forecast students' financial literacy scores and firm's financial distress. In the third Chapter, we innovate the applied machine learning literature by proposing a novel sensitivity analysis for predictions. Finally, in the fourth Chapter, we show how economic intuition can boost the performance of machine learning algorithms. The Dissertation contributes to the literatures on causal and predictive machine learning mainly by: (i) extending the current framework to novel scenarios and applications (Chapter 1 - Chapter 3); (ii) introducing interpretability in the learning models (Chapter 1 - Chapter 2); (iii) developing a novel methodology to assess the robustness of predictions (Chapter 3); (iv) informing the choice of the technique used by specific economic knowledge on the field of investigation (Chapter 4). In the applied Sections of each Chapter, we answer policy relevant questions that pave the way to the usage of machine learning for targeted interventions.

Introduction

In recent years, the ability of machines to solve increasingly more complex tasks has grown exponentially (Sejnowski, 2018). The availability of learning algorithms that deal with tasks such as facial and voice recognition, automatic driving, and fraud detection (among the others) makes the various applications of *artificial intelligence* a hot topic not just in specialized literature but also in other media outlets. The development of *intelligent machines* that can mimic inherently human functions such as learning and problem solving (Russell & Norvig, 2016) is revolutionizing our economy, healthcare, welfare and transports systems, not to mention our cities, houses, jobs, and, ultimately, our lives.

Even if artificial intelligence is often perceived as a relatively recent field of study, some of the most wide spread algorithms in its sub-field of *machine learning*[1] are the improved versions of learning algorithms that were first developed at the beginning of the second half of the twentieth century. In fact, for many decades, computer scientists have been using algorithms that automatically update their course of action to better their performance. Already in the 1950's, Arthur Samuel developed a program to play checkers that improved its performance by learning from its previous moves. The term "machine learning" was coined in that context, and is the field of study that gives computers the ability to learn without

This Chapter extends Bargagli-Stoffi, Niederreiter and Riccaboni (2020).

[1] Artificial intelligence is divided in subfields based on the different tasks being tackled: e.g., *machine learning* deals with the design and implementation of learning algorithms, *natural language processing* deals with speech recognition, understanding and reproduction, *robotics* deals with the design and construction of robots, and so on.

being explicitly programmed, according to Samuel (1959).

Social and health sciences haven't been immune to the large spread of machine learning applications. Starting from the call of Varian (2014) on the potential applications of machine learning in economics, there has been a rise in the usage of machine learning to address issues related to policy prediction and evaluation. Indeed, machine learning provides a set of powerful tools that can be applied both to investigate at a finer and more granular level research questions that were already been raised in previous literature (Cockx et al., 2019; Davis & Heller, 2017; Dubé & Misra, 2017; Kleinberg et al., 2015), and to open new fields of investigation (Kleinberg et al., 2017; Plonsky et al., 2017). The huge impact of these methodologies on the advancement of sciences is proved by the fact that the prestigious scientific journal *Nature* recently kicked off a new collection on the topic of "Machine learning for healthcare"[2].

To better understand how and why these methods are having such a huge impact, let us provide an insight on how machine learning works in practice. Tom M. Mitchell defined the *learning* component of machine learning as follows: "*a computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P** if its performance at tasks in **T**, as measured by **P**, improves with experience **E***" (Mitchell, 1997, p. 2). The **E** component of the algorithm can be thought as a set of data that are used as an input for the machine learning model, while the performance measure **P** is a measure of the quality of model's outputs[3]. Most machine learning methods can be divided in two main branches with respect to the task **T** being addressed: (i) *unsupervised learning* and (ii) *supervised learning* models. Unsupervised learning refers to those techniques used to draw inferences from datasets consisting of input data without labelled outcomes . These algorithms are used to perform tasks such as clustering and pattern mining. Supervised learning refers to the class of algorithms employed to make predictions on la-

[2] This paper's collection can be found on Nature's website at the following address: `https://www.nature.com/collections/zbkpvddmhm`.

[3] The choice of the right performance measure is crucial and is covered in more detail in Section 1.

belled outcome values (i.e., discrete and continuous outcomes)[4]. These techniques use a known dataset with input data and outcome values, referred as training dataset, to learn how to successfully perform predictions on the labelled outcomes. The learned decision rules can then be used to predict unknown outcomes for a new set of observations. Machine learning algorithms provide great added value in such predictive tasks since they are specifically designed for such purposes (Kleinberg et al., 2015). Moreover, the non-parametric nature of machine learning algorithms makes them suited to uncover hidden relationships between the predictors and the outcome variable that could be overlooked by traditional econometric approaches. Indeed, the latter models – e.g., ordinary least squares and logistic regression – are built assuming a set of restrictions on the functional form of the model and on the statistical distributions of the errors to guarantee statistical properties such as estimator unbiasedness and consistency. Machine learning algorithms often relax those assumptions and the functional form is dictated by the data at hand (the phrase "data-driven models", often used to refer to machine learning algorithms, highlights this feature). This characteristic makes machine learning algorithms more *adaptive* and *inductive*, therefore enabling more accurate predictions for future outcome realizations.

In this Dissertation, we deal with a series of applications of machine learning in the fields of social and health science. We introduce a set of novelties in the traditional usage of machine learning algorithms for predictive and causal inference tasks. Before getting into the nuts and bolts of the Dissertation, let us set the stage on machine learning and its applications. In Section 1, we introduce a very general overview on machine learning and a high-level discussion on the most widely used algorithms. Section 2 briefly introduces the main applications of machine learning in social sciences and categorizes them. In Section 3, we draw the outline of the Dissertation and we highlight how each Chapter contributes to the related machine learning literature.

[4] Henceforth, we refer to supervised learning algorithm simply as "machine learning" as it is common usage in the applied machine learning literature.

1 Machine learning

In his famous paper on the difference between model-based and data-driven statistical methodologies Leo Breiman stated, referring to the statistical community, that *"there are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.* [...] *If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a diverse set of tools"* (Breiman et al., 2001, p. 199). In this quote Breiman catches the essence of machine learning algorithms: their ability to capture hidden patterns in the data by directly learning from them, without the restrictions and assumptions of model-based statistical methods.

Machine learning algorithms employ a set of data with input data and outcome values, referred to as training sample, to learn and make predictions (*in-sample* predictions) while another set of data, referred to as test sample, is kept separate to validate the predictions (*out-of-sample* predictions). Training and testing sets are usually built by randomly sampling observations from the initial dataset. In the case of panel data, the testing sample should contain only observations that occurred later in time than the observations used to train the algorithm to avoid the so-called *look-ahead bias*. This ensures that future observations are predicted from past information, not vice versa.

When the dependent variable is categorical (e.g., yes/no, categories 1 to 5, etc) the task of the machine learning algorithm is referred to as a *classification* problem, whereas when the dependent variable is continuous it is referred to as a *regression* problem.

The common denominator of machine learning algorithms is that they take an information set $\mathbf{X}_{N\times P}$ – namely, a matrix of features (also referred to as attributes or predictors) – and map it to an N-dimensional vector of outputs y (also referred to as actual values, dependent variable or response variable), where N is the number of observations $i = 1, \ldots, N$ and P is the number of features. The functional form of this relationship is very flexible and gets updated by evaluating a loss function. The functional form is usually modelled in two steps (Mullainathan & Spiess, 2017):

1. pick the best in-sample loss-minimizing function $f(\cdot)$:

$$argmin \sum_{i=1}^{N} L\big(f(x_i), y_i\big) \quad over \quad f(\cdot) \in F \qquad s.\,t. \qquad R\big(f(\cdot)\big) \leq c \tag{1}$$

   where $\sum_{i=1}^{N} L\big(f(x_i), y_i\big)$ is the in-sample loss functional to be minimized (i.e., the mean-squared-error of prediction), $f(x_i)$ are the predicted (or fitted) values (commonly reported also as $\hat{y}_i$), y_i are the actual values, $f(\cdot) \in F$ is the function class of the machine learning algorithm and $R\big(f(\cdot)\big)$ is the complexity functional that is constrained to be less than a certain value $c \in \mathbb{R}$[5];

2. estimate the optimal level of complexity using empirical tuning through *cross-validation*.

Cross-validation refers to the technique that is used to evaluate predictive models by training them on the training sample and evaluating their performance on the test sample[6]. Then, on the test sample the algorithm's performance is evaluated on how well it has learned to predict the dependent variable y. By construction, many machine learning algorithms tend to perform extremely well on the training data. This phenomenon is commonly referred to as *over-fitting the training data* because the model shows a very high predictive power on the training data and, consequently, a poor fit on the test data. This lack of *generalizability* of the model's prediction from one sample to another can be addressed by penalizing the model's complexity. The choice of a good penalization algorithm is crucial for every machine learning technique to avoid this class of problems.

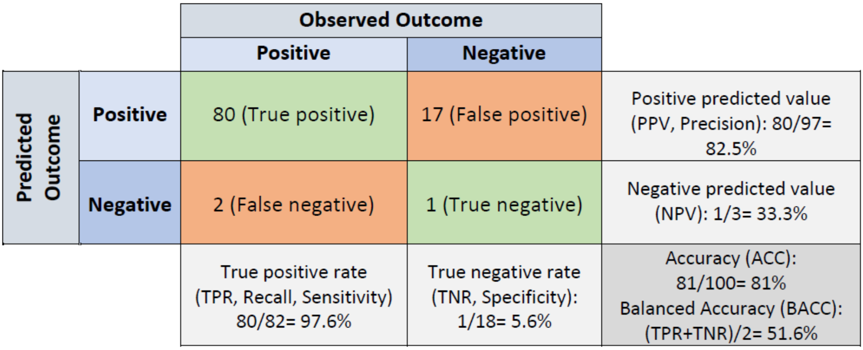
[5] For the sake of simplicity, one can think of this parameter as a *budget constraint* that limits the complexity of the model.

[6] This technique (*hold-out*) can be extended from two to k folds. In k-folds cross-validation, the original dataset is randomly partitioned into k different subsets. The model is constructed on $k-1$ folds and evaluated on 1 fold repeating the procedure until all the k folds are used to evaluate the predictions.

To optimize the complexity of the model, the performance of the machine learning algorithm can be assessed by employing various performance measures on the test sample. It is important for scholars and practitioners to choose the performance measure that best fits the prediction task at hand and the structure of the response variable. In regression tasks different performance measures can be employed. The most common are the mean-squared-error (MSE), the mean-absolute-error (MAE) and the R^2 (and its adjusted version). In classification tasks the most straightforward method is to compare true outcomes with predicted ones via *confusion matrices* from where common evaluation metrics, such as true positive rate (TPR), true negative rate (TNR), and accuracy (ACC), can be easily calculated (see Figure 1). Another popular measure of prediction quality for binary classification tasks (i.e., positive vs. negative response), is the Area Under the receiver operating Curve (AUC) that relates how well the trade-off between the model's TPR and TNR is solved. TPR refers to the proportion of positive cases that are predicted correctly by the model, while TNR refers to the proportion of negative cases that are predicted correctly. Values of AUC range between 0 (total misprediction) and 1 (perfect prediction), where 0.5 indicates that the model has the same prediction power as a random assignment. The choice of the appropriate performance measure is key to communicate the fit of a machine learning model in an informative way.

Consider the example in Figure 1 in which the testing data contains 82 positive outcomes (e.g., firm survival) and 18 negative outcomes (e.g., firm exit), and the algorithm predicts 80 of the positive outcomes correctly but only one of the negative ones. The simple accuracy measure would indicate 81% correct classifications, but the results suggest that the algorithm has not successfully learned how to detect negative outcomes. In such a case, a measure that considers the unbalanced outcomes in the testing set, like balanced accuracy (BACC), defined as $((TPR + TNR/2) = 51.6\%)$, or the F1-score would be more suited. Once the algorithm has been successfully trained and its out-of-sample performance has been properly tested, its decision rules can be applied to predict the outcome of new observations, for which outcome informa-

Figure 1: Exemplary confusion matrix for assessment of classification performance.



		Observed Outcome		
		Positive	Negative	
Predicted Outcome	Positive	80 (True positive)	17 (False positive)	Positive predicted value (PPV, Precision): 80/97= 82.5%
	Negative	2 (False negative)	1 (True negative)	Negative predicted value (NPV): 1/3= 33.3%
		True positive rate (TPR, Recall, Sensitivity) 80/82= 97.6%	True negative rate (TNR, Specificity): 1/18= 5.6%	Accuracy (ACC): 81/100= 81% Balanced Accuracy (BACC): (TPR+TNR)/2= 51.6%

Source Bargagli-Stoffi, Niederreiter, et al. (2020).

tion is not (yet) known.

Choosing a specific machine learning algorithm is crucial since *performance*, *computational scalability* (i.e., the ability of an algorithm to efficiently handle a growing amount of data), and *interpretability* differ widely across available implementations. In this context, easily interpretable algorithms are those that provide comprehensive decision rules from which a user can retrace results (I. Lee & Shin, 2019). Usually highly complex algorithms require a discretionary fine tuning of some model hyper-parameters, more computational resources, and so their decision criteria are less straightforward. Yet, the most complex algorithms do not necessarily deliver the best predictions across applications (Kotthoff, 2016). Therefore, practitioners usually run a *horse race* on multiple algorithms and choose the one that provides the best balance between interpretability and performance on the task at hand[7]. In some learning applications, for which prediction is the sole purpose, different algorithms are combined so that overall predictive performance is maximized. Learning algorithms that are formed by multiple self-contained methods are called

[7] We want to stress that the concept of "best balance between interpretability and performance" varies depending on the task at hand, and it is often based on scholars and practitioners' subjective decisions.

ensemble learners (e.g., the super-learner algorithm by Van der Laan et al., 2007).

Moreover, machine learning algorithms are used by scholars and practitioners to perform predictors selection in high-dimensional settings (e.g., scenarios where the number of predictors is larger than the number of observations: small N large P settings), text analytics, and natural language processing (NLP). The most widely used algorithms to perform the former task are the Least-Absolute-Shrinkage-and-Selection-Operator (LASSO) algorithm (Tibshirani, 1996) and its related versions such as stability selection (Meinshausen & Bühlmann, 2010), C-LASSO (L. Su et al., 2016) and R-LASSO (Belloni, Chernozhukov, Hansen, & Kozbur, 2016). The most popular natural language processing and text analytics machine learning algorithms are Support Vector Machines (Steinwart & Christmann, 2008), Naïve Bayes (Ng & Jordan, 2002), and Artificial Neural Networks (Hassoun et al., 1995).

Even though some algorithms provide reasonably good predictions across applications, it is generally difficult to understand afterwards why a certain observation was assigned to a certain value, or why it was classified in a certain way. This issue relates to a possible trade-off between model interpretability and predictive performance. Indeed, in the presence of high-dimensional non-linear data, less complex algorithms are usually the ones that are easier to interpret, but also the ones associated with the worst performance. In some applications for which prediction is the sole purpose, a deeper analysis is not required, but commonly it is relevant to identify the attributes that mostly influenced the prediction outcome. There are different ways to investigate the decision rules of the algorithms and to get a sense of variables' importance (e.g., tree-based methods such as Random Forest). Papers such as the one by Fisher et al. (2019) provide robust ways to investigate which the variables with the highest predictive power are.

Reviewing machine learning in detail would go beyond the scope of this Introduction, however, in Table 3 we provide a basic intuition of the machine learning methodologies employed in the Dissertation.

Table 1: Machine learning algorithms

Method	Description	Interpretability	Computational Scalability	Used in Chapter(s)
Decision Tree	Decision Trees consist of a sequence of binary decision rules (nodes) on which the tree splits into branches (edges). At each final branch (leaf node) a decision regarding the outcome is estimated. The sequence and definition of nodes is based on minimizing a measure of node purity (e.g., Gini index, or entropy for classification tasks and MSE for regression tasks). Decision Trees are easy to interpret but sensitive to changes in the features, that frequently lower their predictive performance (see also Breiman, 2017).	High	High	2,3,4
Random Forest	Instead of estimating just one Decision Tree, Random Forest re-samples the training set observations to estimate multiple trees. For each tree at each node a set of m (with $m < P$) predictors is chosen randomly from the features space. To obtain the final prediction, the outcomes of all trees are averaged or in classification tasks chosen by majority vote (see also Breiman, 2001).	Medium	Medium	2,3,4,5
Bayesian Additive Regression Trees	Similarly to the Random Forest algorithm, Bayesian Additive Regression Tree is a sum-of-trees ensemble methodology. It relies on a fully Bayesian probability model to obtain its predictions. The Bayesian component of the algorithm is incorporated in a set of three different priors on: (i) the structure of the trees; (ii) the distribution of the outcome in the nodes; (iii) the error variance (see also Chipman et al., 2010).	Medium	Medium-low	2,3,4,5
Super Learner	The idea behind the Super Learner is to combine some candidate algorithms, called learners, to create a new predictive model. The weighted convex combination of the single learner predictions, leads to the final super-learner fit and the optimal combination is obtained through cross-validation (see also Van der Laan et al., 2007)	Low	Low	3

2 Literature on applied machine learning

In the previous Section, we highlighted how the strength of machine learning comes from the fact that it provides a very flexible way to make high-quality predictions, at the expenses of un-verifiability (or, sometimes, absence) of the model underlying assumptions. To put this in a way that is familiar to statisticians, econometricians and quantitative social scientists, we could simplify by saying that machine learning is more about getting a precise estimation of the output of interest, say $\hat{y}$, than a stable, unbiased estimate of the parameters of the model, say $\hat{\boldsymbol{\beta}}$. Hence, the largest part of applications of machine learning in social sciences regards topics where the added value of improved predictions is large (Kleinberg et al., 2015).

As machine learning algorithms are a relatively new set of tools, it is important to try to categorize the various streams of literature of applied machine learning. To do so, we rely on two review papers in this field by Mullainathan and Spiess (2017) and Athey and Imbens (2019). The former paper contributes to a topic-based categorization of the papers, while the latter to a methodological one. Here, we build on the categorization proposed by Mullainathan and Spiess (2017) and we integrate it with the insights provided by Athey and Imbens (2019), with a focus on the field of causal machine learning.

Mullainathan and Spiess (2017) highlight how the largest part of applications of machine learning in social sciences regard four subfields: (i) machine learning for prediction in policy; (ii) machine learning for causal inference; (iii) machine learning for the construction of new data sources; (iv) machine learning to test theories. As this Dissertation is aimed at contributing to the first two subfields, we introduce the related two streams of literature[8].

A complete review of the various applications of machine learning for prediction policy and causal issues would go beyond the scope of the present Introduction. Nevertheless, we want to provide to the reader a

[8] We refer the reader to the original paper by Mullainathan and Spiess (2017) if she wants to deepen into machine learning applications for the construction of new data sources and theory testing.

brief introduction to some of the main scientific works in these fields to better place our contributions in these streams of literature. The first subfield regards the most "straightforward" applications of the predictive power of machine learning. Indeed, as Kleinberg et al. (2015) point out in their work, there are many policy applications in which causality is not central, or even necessary, and scholars need to rely on predictive algorithms. As a toy example, let us imagine a policy-maker facing potentially adverse weather conditions. On the one side, if she wants to assess whether or not a rain-dance would help, she needs to answer to a *causal* question (e.g., "could a rain-dance stop the rain?"). On the other side, if she wants to know whether or not to buy umbrellas for her constituents, she needs to answer a *predictive* question, (e.g., "will it pour with rain?").

The latter question refers to a prediction policy problem and can be answered by using a machine learning model with a set of appropriate predictors. Prediction policy problems of various sorts have been addressed in the literature as machine learning models can be adapted to a wide and heterogeneous set of issues. Kleinberg et al. (2017) use a tree-based algorithm to help judges improve their decisions on bail problems regarding whether or not to release a defendant before the trial. The authors show that machine learning could help policy-makers to either decrease the number of defendants detained or the risks of defendants' recidivism and fail-to-appear, while reducing racial bias. Job market selection has been widely studied through the lens of machine learning. For instance, Kane and Staiger (2008) and Jacob et al. (2018) show how machine learning predictions, based on the observed characteristics of teachers, can predict their teaching quality. More generally, Chalfin et al. (2016) show that machine learning tools can provide a value added when facing human resources decisions. Besides job market selection, machine learning is widely used for the prediction of firms' dynamics[9]. Indeed, Moscatelli et al. (2019) argue that machine learning outperforms standard econometric models when the predictions of firms' distress is (i) based solely on financial accounts data as predictors and, (ii) relies on

[9] We refer to Bargagli-Stoffi, Niederreiter, et al. (2020) for a detailed review of this stream of literature.

a large amount of data. In fact, as these algorithms are *model-free*, they need large datasets (that is the reason why they are sometimes referred to as "data hungry algorithms") to extract the amount of information needed to build precise predictive models. Table 2 depicts several papers in the field of economics, computer science, statistics, business and decision sciences that deal with the issue of predicting firms' bankruptcy or financial distress through machine learning algorithms. The former stream of literature (*bankruptcy prediction*) - that has its foundations in the seminal works of Udo (1993), K. C. Lee et al. (1996), Shin et al. (2005) and Chandra et al. (2009) - compares the binary predictions obtained with machine learning algorithms with the actual realized failure outcomes and uses this information to calibrate the predictive models. The latter stream of literature (*financial distress prediction*) - pioneered by Fantazzini and Figini (2009) - deals with the problem of predicting default probabilities (Moscatelli et al., 2019) or financial constraint scores (Linn & Weagley, 2019). Even if these streams of literature tackle the issue of firms' viability from slightly different perspectives, they train their models on dependent variables that range from firm's bankruptcy (see all the "bankruptcy papers" in Table 2) to firm's default (Behr & Weinblat, 2017; Fantazzini & Figini, 2009; Moscatelli et al., 2019), liquidation (Bonello et al., 2018), firm's political connections (Mazrekay et al., 2018), and financial constraints (C. Hansen et al., 2018; Sun et al., 2017).

Above and beyond applications for prediction in policies, causal problems can be addressed using machine learning as well. Before getting in the nuts and bolts of this stream of literature, it is important to highlight how the usage of machine learning techniques and data science methodologies does not rule out the centrality of the usual identification assumptions that are at the foundations of any causal inference application[10]. Machine learning in this sense, offers a set of tools that widens and sharpens traditional causal inference methodologies, without changing in a substantial way the usual causal framework. Scholars and practitioners eager to use causal machine learning techniques,

[10] For a general review of causal inference in the era of data science we refer the reader to Dominici and Mealli (2020).

should still check for assumptions – such as (among the others) unconfoundedness, positivity, consistency, no-interference (Imbens & Rubin, 2015) – to hold. As the causal machine learning literature is growing at a very fast pace, we highlight some of the main fields of applications of this field to this day. Causal machine learning methodologies have been applied to tackle issues regarding (i) potential outcomes imputation, (ii) doubly robust methods for the estimation of treatment effects, (iii) heterogeneous effects detection and estimation, and (iv) instrumental variables selection[11]. The seminal contributions in the field of causal machine learning are direct applications of the predictive power of machine learning for potential outcomes imputation. In fact, as causal inference is fundamentally a missing data problem (Holland, 1986), machine learning provides a straightforward way to impute the missing potential outcome using the information for the observed potential outcomes for units with the same treatment level as training data. Starting from this idea, Foster et al. (2011) propose to use Random Forest (Breiman, 2001) to perform such an imputation, while Hill (2011) develops a similar method based on Bayesian Additive Regression Trees (Chipman et al., 2010). Following these seminal contributions, various techniques have been proposed to estimate the missing potential outcomes and the propensity score in a flexible way that combines these two estimation problems in doubly robust methods (Chernozhukov et al., 2017; Chernozhukov et al., 2016). Moreover, besides the estimation of the average treatment effect, researchers are interested in investigating how the treatment effects vary across different subpopulations. These subpopulations can be either defined ex-ante (through some intuition on possible sources of effect variation for the phenomenon being studied), or scholars may want to *let the data speak for themselves* to avoid potential cherry-picking in the selection of the subpopulations (Cook et al., 2004). In the latter case, different

[11] Other recent applications of causal machine learning are in panel data scenarios where just one or very few units receive the treatment (Ben-Michael et al., 2018), settings where there is no overlap between treated and control units (Nethery et al., 2019), or in the estimation of causal effects in the presence of interference (Bargagli-Stoffi, Tortu, et al., 2020).

A more detailed set of references regarding this stream of literature will be provided in Chapters 1 and 2.

data-driven techniques have been used for this scope: among the others, Athey and Imbens (2016) develop the *honest causal tree* algorithm building on Classification and Regression Tree (Breiman et al., 1984). Wager and Athey (2018) and Athey et al. (2019) extend this framework by developing the Causal Forest and Generalized Random Forest algorithms, while P. R. Hahn et al. (2020) proposes a Bayesian counterpart of the Generalized Random Forest named Bayesian Causal Forest. These techniques can be adapted for interpretable causal rules detection (K. Lee et al., 2018) and for optimal policy estimation (Athey & Wager, 2017; Kallus, 2018; Z. Zhou et al., 2018). Furthermore, it may be the case that researchers are dealing with multiple instrumental variables issues and they need some guidance regarding which one to select. Belloni et al. (2014), Belloni, Chernozhukov, Hansen, and Kozbur (2016) propose to use LASSO, or ridge regression (C. Hansen & Kozbur, 2014) to perform instrumental variables selection.

3 Dissertation outline

The present Dissertation is divided in two parts. In Part 1, we explore the field of machine learning for causal inference and we introduce two innovative techniques that combine state-of-the-art machine learning algorithms with causal inference methodologies. In the first Chapter, we introduce a novel Bayesian tree-based methodology to draw causal inference on heterogeneous effects in quasi-experimental scenarios. In the second Chapter, we account for possible drawbacks of tree-based methodologies by proposing a composite algorithm with a high level of interpretability and precision. In Part 2, we introduce applications of machine learning predictive power to forecast students' financial literacy scores and firm's financial distress. In the third Chapter, we innovate the applied machine learning literature by proposing a novel sensitivity analysis for predictions. Finally, in the fourth Chapter, we show how economic intuition can boost the performance of machine learning algorithms.

The Dissertation contributes to the literatures on causal and predictive machine learning mainly by: (i) extending the current framework to novel scenarios and applications (Chapter 1 - Chapter 3); (ii) introducing interpretability in the learning models (Chapter 2 - Chapter 4); (iii) developing a novel methodology to assess the robustness of predictions (Chapter 3); (iv) informing the choice of the technique used by specific economic knowledge on the field of investigation (Chapter 4).

Below, we introduce in deeper detail these works, how they contribute and innovate the existing literature, and the main research questions (RQs) that we tackle therein.

3.1 Part 1: Machine learning for causality

Part 1.1: Machine learning for heterogeneous effects

The gold standard for assessing the causal effects of an intervention is to conduct a Randomized Control Trial (RCT). For instance, Duflo et al. (2015) randomly selected Kenyan schools to receive funds to hire additional teachers and then evaluate the effects of this policy on students' performance. However, it is often the case that researchers, for financial and ethical reasons, cannot set up an experiment to assess the impact of a given policy. In such scenarios, researchers must find quasi-experimental set ups that allow them to draw causal conclusion.

In the first Chapter, we develop a novel machine learning algorithm to draw inference on the heterogeneity of causal effects in quasi-experimental set ups where the treatment is not randomly assigned to the treated units. This work innovates the literature on the estimation of heterogeneous causal effects, by providing a tool to discover and estimate heterogeneous effects in the presence of non-compliance between the assignation and the reception of the treatment (these scenarios are sometimes referred as "broken randomizations"). We extend the seminal work by Bargagli-Stoffi and Gnecco (2018), by developing a novel methodology that, on the one side improves over state-of-the-art methodologies such as Causal Trees with Instrumental Variables (Bargagli-Stoffi & Gnecco, 2020) and *Generalized Random Forests* (Athey et al., 2019), and on the other guarantees statistical properties of the conditional estimators such as consistency and asymptotic normality. We show, through Monte

Carlo simulations, that our proposed Bayesian Causal Forest with Instrumental Variable (BCF-IV) algorithm outperforms other machine learning techniques tailored for causal inference in estimating the heterogeneous causal effects.

BCF-IV provides the policy-makers with a relevant tool for real-world evidence on treatment effect variations in observational studies. Such insights on heterogeneous effects are utterly relevant for targeted policies. Indeed, as policy-makers are always facing financial constraints, they may want their policies to just target subsets of the population that really benefit from it (or, in turn, do not target those that do not benefit).

This is particularly true for education policies. As human capital plays a fundamental role in modern societies (Romer, 1989), it is deemed important to understand which the characteristics of students that get the highest return on their performance from additional school transfers are. Hence, the research question the we investigate in the first Chapter is the following:

> **RQ1:** ***What kinds of students most benefit from additional school funding?***

As funding is generally not assigned in a randomized way, traditional causal machine learning techniques may fail in investigating such a research question, providing biased estimates of the true conditional causal effects. BCF-IV could be used for such a task, accounting for broken-randomization, greatly enhancing the effectiveness of school funding. Hence, we use BCF-IV to evaluate the heterogeneity in the effects of a program, promoted by the Flemish Ministry of Education starting from 2002, aimed at providing additional funding for secondary schools which have a high share of disadvantaged students (De Witte et al., 2018). We focus on the effects of additional funding on two outcomes: (i) students' performance and (ii) students' progresses to the following year without retention in their grades. We run our analysis on the universe of pupils (135,682 students) in the first stage of secondary education in the school year 2010/2011 in Flanders.

Part 1.2: Machine learning for interpretable causal rules

A main drawback of machine learning methods in general, and causal machine learning models in particular, is that they may lack interpretability. As causal machine learning is employed more and more often to aid policy-makers in high-stakes scenarios, it is deemed important to provide information on simple causal rules (e.g., in the form of a collection of heterogeneous subgroups to target) that can be exploited to improve policy effectiveness and to ensure that stakeholders and policy-makers understand (and, in turn, trust) the functionality of these models. In the second Chapter, we propose a new causal rule ensemble (CRE) method that has two main features: 1) ensures interpretability of the causal rules; and 2) provides a level of statistical precision of the estimated conditional causal effects for each of the newly discovered rules that is comparable to the most accurate causal machine learning algorithms.

Besides providing methodological innovations, CRE accommodates for central problems in the discovery of interpretable heterogeneous effects. In fact, tree-based methods provide the highest level of interpretability but have several limitations: (i) they are unstable; (ii) they allow for the discovery of a limited set of subgroups (i.e., subgroups that can be detected through a binary Decision Tree); (iii) they may fail to distinguish between effect modifiers and confounders.

To adjust for these limitations and develop an interpretable algorithm, we use a sample-splitting approach that splits the total sample into two smaller subsamples: the discovery subsample and inference subsample. In the discovery subsample, we generate a large set of causal rules through state-of-the-art causal machine learning methodologies, and then select a few of them through regularization. In the inference subsample, we estimate the causal treatment effect for each rule and its confidence interval. We provide theoretical results that guarantee consistency of the estimated causal effects for the newly discovered rules. Furthermore, via Monte Carlo simulations, we show that the causal rule ensemble method has an excellent performance with respect to its ability to discover the rules and then accurately estimate the causal effects, under different scenarios.

Effect variation discovery is particularly important when we deal with health-related issues and it is at the very foundations of the recently established field of *personalized medicine*. In the application section of the Chapter, we use the proposed method to the study of the effects of long-term exposure to air pollution on the 5-year mortality rate of the New England Medicare population in United States finding the subgroups of people who are more vulnerable. While in the first Chapter our goal is to estimate the subpopulation that benefits the most from the intervention, here we want to discover the people that are more at risk of developing diseases as a consequence of exposure to air pollution. Hence, our research question is the following:

> **RQ2:** ***What characteristics make people most vulnerable to the negative effects of higher levels of pollution?***

To investigate this research question, we use data on Medicare beneficiaries in New England regions of the United States between 2000 and 2006 and the pollution levels for the same regions and time span.

3.2 Part 2: Machine learning for prediction

Part 2.1: Using machine learning for robust predictions in a novel application

In the third Chapter, we propose a general algorithm to assess how the omission of an unobserved variable affects the predictions and the performance of a machine learning model. We argue that there is a need for specific techniques which can assess, what we call, the *sensitivity* of forecasts to variations in the set of predictors used to train the model. Indeed, we claim that a desirable property of any (well-performing) predictive model is not to show extreme variations in its prediction and accuracy following the inclusion of a novel variable in the model. We claim that assessing the stability of an algorithm is central in any high-stakes prediction.

The proposed methodology extends the usage of machine learning from pointwise predictions to inference and robustness analysis. In the application, we show how our framework can be applied to data with inherent uncertainty, as school outcomes. First, using Bayesian Additive Regression Trees (Chipman et al., 2010), we predict students' financial literacy scores (FLS) for a pool of students with missing scores. We apply our proposed framework to the 2015 data of the OECD's Program for International Student Assessment (PISA) for Belgium. Belgium provides a very interesting case study as all regions of the country participated in the general assessment of PISA, while only selected regions participated in the financial literacy assessment. This structure of the data allows us to use a common set of predictors for all regions from the general assessment to predict financial literacy outcomes for students in the regions that did not participate in this part of PISA. Then, we assess the sensitivity of predictions by comparing models with and without a synthetic predictor which highly correlates with the outcome variable but is uncorrelated to the observed set of predictors. The methodology that we develop provides an answer to the following research question:

> **RQ3:** ***How sensitive are the predictions of student financial literacy scores to a potentially unobserved predictor with high explanatory power?***

Our proposed framework has applications beyond the scope of financial literacy. From a policy perspective, it creates the opportunity to predict missing observations in large administrative datasets in service of targeted policies, and it explores how robust machine learning predictions are.

Part 2.2: Improving machine learning predictions through economic intuition

In the fourth and last Chapter, we exploit machine learning techniques in a prediction policy setting. Our task is to improve current methodologies in predicting the risk of failure of firms and to provide a novel definition of *zombie firms*. This Chapter contributes both to the policy prediction literature, to the literature on firms' distress, bankruptcy prediction and *zombie firms* by introducing for the first time in this setting a machine learning methodology that outperforms all the state-of-the-art methods employed thus far in the literature (see Table 2 on how our contribution compares to the literature).

The estimation of default probabilities, financial distress and the predictions firms' bankruptcies based on balance sheet data and other sources of information on firms' viability is a highly relevant topic for regulatory authorities, financial institutions, and banks. In fact, regulatory agencies often evaluate the ability of banks to assess enterprises viability as this affects their capacity of best allocating financial resources and, in turn, their financial stability. Hence, a higher predictive power of machine learning algorithms can boost targeted financing policies that lead to safer allocation of credit.

We show that, on the one side, informed choices of the predictors can boost the performance of the machine learning model, and on the other side economic intuition can guide researchers in the choice of the machine learning algorithm to be used for the data analysis. In fact, we find that machine learning methodology that incorporates the information on missing data into its predictive model – i.e., the Bayesian Additive Regression Trees with Missing Incorporated in Attributes algorithm by Kapelner and Bleich (2015) – can lead to staggering increases in the predictive performances when the predictors are missing-non-at-random (MNAR) and their missingness patterns are correlated with the outcome. We argue that often times the decision not to release financial account information is driven by firm's financial distress. Hence, the research question that we investigate is the following:

> **RQ4:** ***Can economic intuition boost the machine learning predictions of firms' distress and provide insight on non-viable firms that do, however, still survive?***

We start by training our algorithm on about 305,000 firms active in Italy in the period 2008-2017. We compare our technique to the use of previous financial indicators, including Z-scores and distance-to-default, more traditional econometric techniques, and other widely used machine learning algorithms. Moreover, we discuss how machine learning helps in identifying *zombie firms*, which we define as firms that persist in status of a high risk of failure. More generally, we highlight how statistical learning may help in the design of target-specific and evidence-based policies in presence of market selection failures, for example in the design of optimal bankruptcy laws.

Table 2: Machine learning literature on firms' failure and financial distress

Author, Year	Domain	Output	Country, time	Dataset size	SL-method	Attributes	GOF
Alaka et al. (2018)	CS	Bankruptcy	UK (2001-2015)	30,000	NN	5	88% (AUC)
Barboza et al. (2017)	CS	Bankruptcy	USA (1985-2014)	10,000	SVM, RF, BO, BA	11	93% (AUC)
Bargagli-Stoffi, Riccaboni, et al. (2020)	**ECON**	**Fin. distress**	**ITA (2008-2017)**	**305,000**	**BART-MIA**	**46**	**97%(AUC)** **63% (F1-score)**
Behr and Weinblat (2017)	ECON	Bankruptcy	INT (2010-2011)	945,062	DT, RF	20	85% (AUC)
Bonello et al. (2018)	ECON	Fin. distress	USA (1996-2016)	1,848	NB, DT, NN	96	78% (ACC)
Brédart (2014)	BMA	Bankruptcy	BEL (2002-2012)	3,728	NN	3	81%(ACC)
Chandra et al. (2009)	CS	Bankruptcy	USA (2000)	240	DT	24	75%(ACC)
Cleofas-Sánchez et al. (2016)	CS	Fin. distress	INT (2007)	240-8,200	SVM, NN, LR	12-30	78% (ACC)
Danenas and Garsva (2015)	CS	Fin. distress	USA (1999-2007)	21,487	SVM, NN, LR	51	93% (ACC)
Fantazzini and Figini (2009)	STAT	Fin. distress	DEU (1996-2004)	1,003	SRF	16	93% (ACC)
C. Hansen et al. (2018)	ECON	Fin. distress	DNK (2013-2016)	278,047	CNN, RNN	50	84% (AUC)
Heo and Yang (2014)	CS	Bankruptcy	KOR (2008-2012)	30,000	ADA	12	94% (ACC)
Hosaka (2019)	CS	Bankruptcy	JPN (2002-2016)	2,703	CNN	14	18% (F1-score)
S. Y. Kim and Upneja (2014)	CS	Bankruptcy	KOR (1988-2010)	10,000	ADA, DT	30	95% (ACC)
K. C. Lee et al. (1996)	BMA	Bankruptcy	KOR (1979-1992)	166	NN	57	82% (ACC)
Liang et al. (2016)	ECON	Bankruptcy	TWN (1999-2009)	480	SVM, KNN	190	82% (ACC)
Linn and Weagley (2019)	ECON	Fin. distress	INT (1997-2015)	48,512	DRF	16	15% (R^2)
Moscatelli et al. (2019)	ECON	Fin. distress	ITA (2011-2017)	250,000	RF	24	84%(AUC)
Shin et al. (2005)	CS	Bankruptcy	KOR (1996-1999)	1,160	SVM	52	77%(ACC)
Sun and Li (2011)	CS	Bankruptcy	CHN	270	CBR, KNN	5	79% (ACC)
Sun et al. (2017)	BMA	Fin. distress	CHN (2005-2012)	932	ADA, SVM	13	87%(ACC)
Tsai and Wu (2008)	CS	Bankruptcy	INT	690-1,000	NN	14-20	79-97%(ACC)
Tsai et al. (2014)	CS	Bankruptcy	TWN	440	ANN, SVM	95	86% (ACC)
G. Wang et al. (2014)	CS	Bankruptcy	POL (1997-2001)	240	DT, NN, NB	30	82% (ACC)
Udo (1993)	CS	Bankruptcy	KOR (1996-2016)	300	NN	16	91% (ACC)
Zieba et al. (2016)	CS	Bankruptcy	POL (2000-2013)	10,700	BO	64	95% (AUC)

Abbreviations used - Domain: ECON: Economics, CS: Computer Science, BMA: Business, Management, Accounting, STAT: Statistics. Country: BEL: Belgium, ITA: Italy, DEU: Germany, INT: International, KOR: Korea, USA: United states of America, TWN: Taiwan, CHN: China, UK: United Kingdom, POL: Poland. Primary SL-method: ADA: AdaBoost, ANN: Artificial Neural Network, CNN: Convolutional Neural Network, NN: Neural Network, GTB: Gradient Tree Boosting, RF: Random Forest, DRF: Decision Random Forest, SRF: Survival Random Forest, DT: Decision Tree, SVM: Support Vector Machine, NB: Naive Bayes, BO: Boosting, BA: Bagging, KNN: K-Nearest Neighbour, BART: Bayesian Additive Regression Tree, DT: Decision Tree, LR: Logistic Regression. Rate: ACC: Accuracy, AUC: Area under the receiver operating curve. The year was not reported when it was not possible to recover this information from the papers. Source: Bargagli-Stoffi, Niederreiter, et al. (2020).

Table 3: Dissertation outline

	Chapter	**Topic**	**Method**	**Application**	**Research Question**
	Introduction	General overview and introduction			
Part 1: Machine learning for causality	Chapter 1	Heterogeneous effects in the presence of imperfect compliance	Bayesian Causal Forest with Instrumental Variables	Funding effectiveness on students' performances in Flanders	What kinds of students most benefit from additional school funding?
	Chapter 2	Interpretable causal rules in observational studies	Causal Rule Ensemble	Effects of pollution on health outcomes in New England (US)	What characteristics make people most vulnerable to the negative effects of higher levels of pollution?
Part 2: Machine learning for prediction	Chapter 3	Understanding how robust are machine learning predictions to unobserved variables	Sensitivity analysis for Bayesian Additive Regression Trees and Random Forest predictions	Predicting non-observed students' financial literacy scores in Belgium	How sensitive are the predictions of student financial literacy scores to a potentially unobserved predictor with high explanatory power?
	Chapter 4	Improving machine learning predictions through economic intuition	Bayesian Additive Regression Trees with Missing Incorporated in Attributes	Forecasting firms' distress in Italy after the 2008 crisis	Can economic intuition boost the machine learning predictions of firms' distress and provide insight on non-viable firms that do, however, still survive?
	Conclusions	General conclusions and discussion			

Part 1
Machine learning for causality

Chapter 1

Heterogeneous causal effects with imperfect compliance

In social and health sciences the largest part of scientific research questions deals with inferring a causal relationship (e.g., evaluating the impact of a policy, the effects of a drug, the returns from a marketing or business strategy and so on). Following the growing availability of large datasets, the necessity to deal with problems connected with potentially heterogeneous treatment effects is stronger than what was observed in the past. The availability of large datasets makes it possible to investigate and, in turn, customize the causal effect estimates for population subsets as well as for individuals (Athey, 2019). In this scenario, machine learning techniques are increasingly used to address causal inference tasks and, in particular, to estimate heterogeneous causal effects (Athey & Imbens, 2016; Foster et al., 2011; Green & Kern, 2012; P. R. Hahn et al., 2020; Hill, 2011; Lechner, 2019; K. Lee et al., 2018; X. Su et al., 2012; Wager & Athey, 2018). A growing literature seeks to apply supervised machine

This chapter is based on Bargagli-Stoffi, De Witte and Gnecco (2019), as well as Bargagli-Stoffi and Gnecco (2018, published in *Proceeding of the 5th IEEE DSAA Conference*) and Bargagli-Stoffi and Gnecco (2020, published in *International Journal of Data Science and Analytics*).

learning techniques to the problem of estimating heterogeneous treatment effects. In their seminal contributions, Hill (2011) and Foster et al. (2011) propose to directly apply machine learning algorithms to estimate the unit level causal effect as a function of the units' attributes. In other contributions, machine learning algorithms are adapted to estimate the heterogeneous causal effects (Athey & Imbens, 2016; Athey et al., 2019; P. R. Hahn et al., 2020; Lechner, 2019; Wager & Athey, 2018). However, most of these techniques are tailored for causal inference in settings where the treatment is randomly assigned to the units and do not address imperfect compliance issues. Nevertheless, in the real world, the implementation of policies or interventions often results in imperfect compliance, which makes the policy evaluation complicated. Imperfect compliance may arise in observational studies where the assignment to the treatment can be different from the receipt of the treatment (e.g., individuals are randomly assigned to a treatment, but not all the units that are assigned to it actually receive it). Recently, some machine learning algorithms have been proposed to deal with imperfect compliance (Athey et al., 2019; Bargagli-Stoffi & Gnecco, 2020; Hartford et al., 2016; Johnson et al., 2019; G. Wang et al., 2018). However, these methods exhibit three principal limitations: (i) random forest-based algorithms for causal inference require large samples to converge to a good asymptotic behaviour for the estimation of causal effects, as shown in P. R. Hahn, Dorie, et al. (2018) and Wendling et al. (2018); (ii) deep learning-based algorithms lack interpretability of the machine learning black-box which can expose them to critiques, especially in the context of social sciences; (iii) the algorithms proposed by G. Wang et al. (2018) and Bargagli-Stoffi and Gnecco (2020), Bargagli-Stoffi and Gnecco (2018) are based on single learning that perform worse as compared to multiple learning (i.e., ensemble methods)[1]. Moreover, in machine learning applications, inference and uncertainty quantification are of secondary importance after predictive performance. However, in policy decision settings it is crucial to know the credibility and variance of the counterfactual predictions

[1] Ensemble methods have extensively been shown to outperform single learning algorithms in prediction tasks (Van der Laan et al., 2007).

(Athey et al., 2019).

To address and accommodate these shortcomings this Chapter innovates the literature in both a methodological and an empirical perspective. First, we develop a machine learning algorithm tailored to draw causal inference in situations where the assignment mechanism is irregular, namely the assignment depends on the observed and unobserved potential outcomes (Imbens & Rubin, 2015). This methodology contributes to the increasing use of machine learning techniques to draw causal inference (Athey & Imbens, 2017). In particular, we propose to modify a machine learning technique, namely, Bayesian Causal Forests, developed for causal inference goals (P. R. Hahn et al., 2020) to fit an instrumental variable setting (Angrist et al., 1996). The proposed method, Bayesian Instrumental Variable Causal Forest (BCF-IV), is an ensemble semi-parametric Bayesian regression model that directly builds on the Bayesian Additive Regression Trees (BART) algorithm (Chipman et al., 2010). BART is an ensemble-of-trees approach to nonparametric regression (Starling et al., 2018), which is in turn a *refined* version of the random forest algorithm (Breiman, 2001): BART obtains more precise estimates both in non-causal inference scenarios (Chipman et al., 2010; Hernández et al., 2018; Linero, 2018; Linero & Yang, 2018; Murray, 2017; Starling et al., 2018) and in causal inference settings (P. R. Hahn, Dorie, et al., 2018; P. R. Hahn et al., 2020; Hill, 2011; Logan et al., 2019) by employing a full set of prior distributions on the depth of the trees, on the noise, and on the outcome in their nodes.

Second, we evaluate the fit of the proposed algorithm by comparing it with two alternative machine learning methods explicitly developed to draw causal inference in the presence of irregular assignment mechanisms: namely, the Generalized Random Forests (GRF) algorithm (Athey et al., 2019) and the Honest Causal Trees with Instrumental Variables (HCT-IV) algorithm (Bargagli-Stoffi & Gnecco, 2020). Using Monte Carlo simulations, we evaluate each algorithm with respect to three dimensions: (i) the choice of the correct source of heterogeneity (i.e., the choice of the right splitting variable); (ii) the choice of the correct cutoff, in the case of a continuous splitting variable, and (iii) the estimation of the heterogeneous causal effects. These dimensions are consistent with recent evaluations of various machine learning methods for causal inference that highlight the excellence of Bayesian algorithms for causal inference (P. R. Hahn, Dorie, et al., 2018; Wendling et al., 2018). We show that for each dimension, BCF-IV outperforms both GRF and HCT-IV in small samples and converges to an optimal asymptotic behaviour.

Third, we show how the proposed algorithm can be used for targeted policies, which are increasingly relevant as the call for personalized interventions has unfurled in all social sciences, especially in economics, health and management sciences (Athey & Imbens, 2017). The main objective of targeted policies studies is to inform policy-makers about the best allocation of treatments to individuals or sub-populations (Kitagawa & Tetenov, 2018). The idea behind these policies is to target those observations that benefit the most from a certain intervention in order to get either of two possible welfare gains: (i) reducing the costs of an intervention with constant effect sizes, or (ii) increasing the intervention effects for given costs (Kleinberg et al., 2017).

Fourth, in an empirical application, BCF-IV is used for the evaluation of an educational policy. The evaluation of educational policies is a promising field for the discovery of heterogeneous causal effects and, in turn, targeted policies. This is due to at least two factors: (i) in the education context, there is a clear source of heterogeneity given by the disparate profiles of schools and students; and (ii) it is possible to gather large (administrative) datasets. In a similar framework, machine learning provides a tailored, data-driven tool for the evaluation of the heterogeneity in the causal effects, and, consequently, the implementation of targeted policies. In particular, we evaluate the effects of additional resources for disadvantaged students on students' performance in a fuzzy Regression Discontinuity Design (RDD) scenario (J. Hahn et al., 2001)[2]. By using a unique administrative dataset, we employ BCF-IV to evaluate the heterogeneity in the effects of the 'Equal Educational Opportunities Program' promoted by the Flemish Ministry of Education starting from 2002. The program is aimed at providing additional funding for sec-

[2] The fuzzy RDD that we implement in this Chapter is described in detail in Section 3.

ondary schools with high share of disadvantaged students (De Witte et al., 2018). We focus on the effects of additional funding on two outcomes: (i) students' performance, namely if a student gets the most favorable outcome (*A-certificate*); and (ii) students' progresses to the following year without retention in their grades. The Flemish Ministry of Education provided us with data on the universe of pupils in the first stage of secondary education in the school year 2010/2011 with a total of 135,682 students. We obtained data on student level and school level characteristics. Moreover, this setting provides us with a quasi-experimental identification strategy since the additional funding is provided to schools based on being above or below an exogenously set threshold regarding the proportion of disadvantaged students. There is also a second, exogenously set, eligibility criterion stating that schools have to generate a minimum number of teaching hours. This provides us with an imperfect compliance setting – as not all the schools fulfill both the criteria – in which we are able to exploit a fuzzy regression discontinuity design to draw causal effects.

The results of our empirical application suggest that, although the effects of additional funding on the overall population of students are found to be not statistically significant[3], there is appreciable heterogeneity in the causal effects: the effects on students' progress are positive and significant if we focus on the sub-population of students in schools with less senior principals (namely, principals with less than 25 years of experience) and younger principals (namely, principals younger than 55 years). These results can advise policy-makers in multiple ways: the heterogeneous drivers could, on the one hand, help them enhancing the policy effectiveness by targeting just the schools with the highest shares of pupils that benefit the most from additional funding. On the other hand, policy-makers could investigate more in depth the reason why some schools do not benefit from the policy and ultimately provide additional tools to these schools to enhance the policy outcomes.

The methodology proposed in this Chapter can be more widely ap-

[3] This is in line with further research of additional funding on school level outcomes (De Witte et al., 2018).

plied to evaluations of the heterogeneous impact of an intervention in the presence of an irregular assignment mechanism in social and biomedical sciences.

The remainder of this Chapter is organized as follows: in Section 1 we provide a general overview on causal inference and the applied machine learning frameworks and we introduce our algorithm. In Section 2 we compare the performance of our algorithm with the performance of other methods already established in the literature. In Section 3 we depict the usage of our algorithm in an educational scenario to evaluate the heterogeneous causal effects of additional funding to schools. Section 4 discusses the results and highlights the further applications of heterogeneous causal effects discovery and targeted policies in education as well as in social and biomedical sciences.

1 Bayesian causal forest with instrumental variable

1.1 Notation

This Chapter contributes to the literature by establishing a novel, combined machine learning approach for the estimation of conditional causal effects in the presence of an irregular assignment mechanism.

We follow the standard notation of the Rubin's causal model (Imbens & Rubin, 2015; Rubin, 1974, 1978). Given a set of N units, indexed by $i = 1, ..., N$, we denote with Y_i a generic outcome variable, with W_i a binary treatment indicator and, with $\mathbf{X}$ a $N \times P$ matrix of P control variables. Given the Stable Unit Treatment Value Assumption (SUTVA), that excludes interference between the treatment assigned to one unit and the potential outcomes of another (Imbens & Rubin, 2015), we can postulate the existence of a pair of potential outcomes: $Y_i(W_i)$. Specifically, the potential outcome for a unit i if assigned to the treatment is $Y_i(W_i = 1) = Y_i(1)$, and the potential outcome if assigned to the control is $Y_i(W_i = 0) = Y_i(0)$. We cannot observe for the same unit both the potential outcomes at the same time. However, we observe the potential outcome that corresponds to the assigned treatment: $Y_i^{obs} = Y_i(1)W_i + Y_i(0)(1 - W_i)$.

In order to draw proper causal inference in observational studies researchers need to assume *strong ignorability* to hold. This assumption states that:

$$Y_i(W_i) \perp\!\!\!\perp W_i | \mathbf{X}_i, \tag{1.1}$$

and

$$0 < Pr(W_i = 1 | \mathbf{X}_i = x) < 1 \;\forall\, x \in \mathbb{X}, \tag{1.2}$$

where $\mathbb{X}$ is the features space. The first assumption (unconfoundedness) rules out the presence of unmeasured confounders while the second condition needs to be invoked to be able to estimate the unbiased treatment effect on all the support of the covariates space. If these two conditions hold, we are in the presence of the so-called *regular assignment mechanism*. In such a scenario the Average Treatment Effect (ATE) can be expressed as:

$$\tau = \mathbb{E}[Y_i^{obs} | W_i = 1] - \mathbb{E}[Y_i^{obs} | W_i = 0], \tag{1.3}$$

and one can define, following Athey and Imbens (2016), the Conditional Average Treatment Effect (CATE) simply as:

$$\tau(x) = \mathbb{E}[Y_i^{obs} | W_i = 1, \mathbf{X}_i = x] - \mathbb{E}[Y_i^{obs} | W_i = 0, \mathbf{X}_i = x]. \tag{1.4}$$

CATE is central for targeted policies as it enables the researcher to investigate the heterogeneity in causal effects. For instance, we may be interested in assessing how the effects of an intervention vary within different sub-populations.

In observational studies, the assignment to the treatment may be different from the reception of the treatment. In these scenarios, where one allows for non-compliance between the treatment assigned and the treatment received, one can assume that the assignment is unconfounded, wherein the receipt is confounded (Angrist et al., 1996). In such cases, one can rely on an instrumental variable (IV), Z_i, to draw proper causal inference[4]. Z_i can be thought as a randomized assignment to the treatment, that affects the receipt of the treatment W_i, without directly affecting the outcome Y_i (*exclusion restriction*). Thus, one can then express the treatment received as a function of the treatment assigned: $W_i(Z_i)$.

If the classical four IV assumptions[5] (Angrist et al., 1996) hold, one can get the causal effect of the treatment on the sub-population of compliers, the so-called Complier Average Causal Effect (CACE), that is:

$$\tau^{cace} = ITT_{Y,C} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]} = \frac{ITT_Y}{\pi_C}. \quad (1.5)$$

CACE is also sometimes referred to as LATE (Local Average Treatment Effects) and represents the estimate of causal effect of the assignment to treatment on the principal outcome, Y_i, for the subpopulation of compliers (Imbens & Rubin, 2015). In this Chapter, we introduce and consider the following conditional version of CATE. The conditional CACE, $\tau^{cace}(x)$, can be thought as the CACE for a sub-population of observations defined by a vector of characteristics x:

$$\begin{aligned} \tau^{cace}(x) &= ITT_{Y,C}(x) = \frac{\mathbb{E}[Y_i|Z_i = 1, \mathbf{X}_i = x] - \mathbb{E}[Y_i|Z_i = 0, \mathbf{X}_i = x]}{\mathbb{E}[W_i|Z_i = 1, \mathbf{X}_i = x] - \mathbb{E}[W_i|Z_i = 0, \mathbf{X}_i = x]} \\ &= \frac{ITT_Y(x)}{\pi_C(x)}. \end{aligned} \quad (1.6)$$

1.2 Estimating conditional causal effects with machine learning

In recent years, various algorithms have been proposed to estimate conditional causal effects (i.e, CATE and $\tau^{cace}(x)$). Most algorithms focus on the estimation of CATE (Athey & Imbens, 2016; Green & Kern, 2012; P. R. Hahn et al., 2020; Hill, 2011; Lechner, 2019; K. Lee et al., 2018; X. Su et al.,

[4] Throughout this Section and throughout the Chapter we assume the instrumental variable to be binary but, one could relax this assumption. However, as there are currently no studies that develop machine learning algorithms for the estimation of heterogeneous causal effects with a continuous treatment variable, we leave the investigation of these algorithms to further research.

[5] See Appendix 5 for a detailed discussion of the four assumptions and how they are assumed to hold in our application reported in Section 3.

2012; Wager & Athey, 2018) while just a few (Athey et al., 2019; Bargagli-Stoffi & Gnecco, 2020; Hartford et al., 2016; G. Wang et al., 2018) focus on the estimation of $\tau^{cace}(x)$. In this Chapter, we propose an algorithm for the estimation of CATE in an irregular assignment mechanism scenario. In particular, we adapt the Bayesian Causal Forest (BCF) algorithm (P. R. Hahn et al., 2020) for such a task. BCF was originally proposed for regular assignment mechanisms. This algorithm builds on the Bayesian Additive Regression Trees (BART) algorithm (Chipman et al., 2010) which is a Bayesian version of an ensemble of Classification and Regression Trees (CART) (Breiman et al., 1984)[6].

CART is a widely used algorithm for the construction of binary trees (namely, trees where each node is splitted into only two branches). A binary tree is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample and proceeding with the splits to the final nodes (leaves). Figure 2 illustrates how the binary partitioning works in practice in a simple case with just two regressors $x_1 \in [0,1]$ and $x_2 \in [0,1]$.

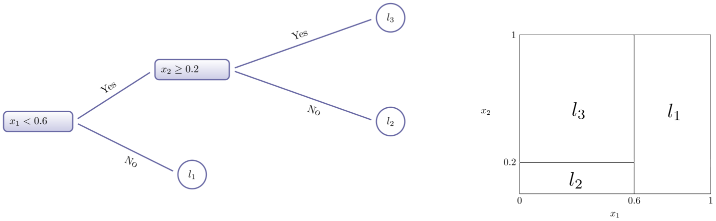


Figure 2: An example binary tree. The internal nodes are labelled by their splitting rules and the terminal nodes labelled with the corresponding parameters l_i.
(Right) The corresponding partition of the sample space.

Binary trees are named *classification trees* when the outcome variable can take a discrete set of values, and *regression trees* when the outcome variable takes continuous values. The CART algorithm associates, for

[6] Chipman et al. (2010) highlight how their algorithm is different from other ensemble methods such as the Random Forest algorithm (Breiman, 2001).

every individual belonging to a partition of the feature space, a conditional prediction for the outcome variable. The task of a binary tree is to estimate the conditional expectation of the observed outcome, on the basis of the information on features and outcomes for units in the training sample, and to compare the resulting estimates on a test sample to tune the complexity of the tree, in order to minimize the "error"[7] between the true and estimated values of $Y_i(x)$ within each partition.

The accuracy of the predictions of binary trees, $\hat{Y}_i(x)$, can be dramatically improved by iteratively constructing the trees. A Random Forest (RF) consists in an ensemble of trees, where each tree is constructed by randomly sampling the observations and randomly drawing the covariates (predictors, in the machine learning literature) that are used to build each tree (Breiman, 2001). One of the main problems of RFs is that they tend to "overfit" the data on which they are trained. "Overfitting" leads to a scarce generalizability of the predictions on samples different from the training set. In order to avoid this, Bayesian Additive Regression Trees were proposed by Chipman et al. (2010).

BART, as well as BCF, are "refined versions" of the RF algorithm. BART is, as the RF, a sum-of-trees ensemble algorithm, but its estimation approach used to obtain the values of $Y_i(x)$ relies on a fully Bayesian probability model (Kapelner & Bleich, 2015). In particular, the BART model can be expressed as:

$$Y_i = f(\mathbf{X}_i) + \epsilon_i \approx \mathbb{T}_1(\mathbf{X}_i) + ... + \mathbb{T}_q(\mathbf{X}_i) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad (1.7)$$

where each of the q distinct binary trees is denoted by $\mathbb{T}$[8].

The Bayesian component of the algorithm is incorporated in a set of three different priors on: (i) the structure of the trees (this prior is aimed at limiting the complexity of any single tree $\mathbb{T}$ and works as a regularization device); (ii) the probability distribution of data in the nodes (this prior is aimed at shrinking the node predictions towards the center of the distribution of the response variable Y_i); (iii) the error variance σ^2 (which

[7] There are various "error" measures used to optimize binary trees. The most widely used are the mean-squared-error for regression trees and the entropy or the Gini index for classification trees.

[8] $\mathbb{T}$ represents the entire tree: its structure, its nodes and its leaves (terminal nodes).

bounds away σ^2 from very small values that would lead the algorithm to overfit the training data)[9]. The aim of these priors is to "regularize" the algorithm, preventing single trees to dominate the overall fit of the model (Kapelner & Bleich, 2015). Moreover, BART allows the researcher to tune the variables' importance by departing from the original formulation of the Random Forest algorithm where each variable is equally likely to be chosen from a discrete uniform distribution (i.e., with probability $\frac{1}{p}$) to build a single tree learner. These Bayesian tools give researchers the possibility to mitigate the "overfitting" problem of RFs and to tune the algorithm with prior knowledge. Given these characteristics BART has shown particular flexibility and an excellent performance in both prediction tasks (Hernández et al., 2018; Linero, 2018; Linero & Yang, 2018; Murray, 2017; Starling et al., 2018) and in causal inference tasks (P. R. Hahn et al., 2020; Hill, 2011; Logan et al., 2019; Nethery et al., 2019).

Thus far, the algorithms that we discussed are tailored to find heterogeneity in the response variable $Y_i(x)$, but are not developed to estimate the heterogeneity in the causal effects. The BCF algorithm proposed by P. R. Hahn et al. (2020) is a semi-parametric Bayesian regression model that directly builds on BART. It, however, introduces some significant changes in order to estimate heterogeneous treatment effects in regular assignment mechanisms (even in the presence of strong confounding). The principal novelties of this model are the expression of the conditional mean of the response variable as a sum of two functions and the introduction, in the BART model specification for causal inference, of an estimate of the propensity score, $E[W_i = 1|\mathbf{X}_i = x] = \pi(x)$, in order to improve the estimation of heterogeneous treatment effects[10]. As depicted

[9] The choice of the priors, and the derivation of the posterior distributions, is discussed in depth by Chipman et al. (2010) and Kapelner and Bleich (2015). Namely, (i) the prior on the probability that a node will split at depth k is $\beta(1+k)^{-\eta}$ where $\beta \in (0,1), \eta \in [0,\infty)$ (these hyper-parameters are generally chosen to be $\eta = 2$ and $\beta = 0.95$); (ii) the prior on the probability distribution in the nodes is a normal distribution with zero mean: $\mathcal{N}(0, \sigma_q^2)$ where $\sigma_q = \sigma_0/\sqrt{q}$ and σ_0 can be used to calibrate the plausible range of the regression function; (iii) the prior on the error variance is $\sigma^2 \sim InvGamma(v/2, v\lambda/2)$ where λ is determined from the data in a way that the BART will improve 90% of the times the RMSE of an OLS model.

[10] It is important to highlight that the propensity score is not used to estimate the causal effects but to moderate the distortive effects in treatment heterogeneity discovery due to

from the results of the Atlantic Causal Inference Conference (ACIC) competition in 2016 and 2017, reported by P. R. Hahn, Dorie, et al. (2018), it was observed that BCF performs dramatically better than other machine learning algorithms for causal inference in the presence of randomized and regular assignment mechanisms.

1.3 Extending BCF to an IV scenario

Bargagli-Stoffi and Gnecco (2020) show that a naïve application of methods developed for the estimation of heterogeneous causal effects in randomized or regular assignment mechanisms would introduce a large bias in the estimation of the heterogeneous causal effects in imperfect compliance settings. Moreover, the authors show that this would lead to very imprecise heterogeneous causal effects estimators. This reason drives the need for a new algorithm, the Bayesian Instrumental Variable Causal Forest (BCF-IV), tailored for causal inference on heterogeneous effects in the presence of irregular mechanisms.

The BCF-IV algorithm is constructed in two steps:

1. Discovering heterogeneity for the conditional intention-to-treat ($ITT_Y(x)$);
2. Estimation of the conditional CACE ($\tau^{cace}(x)$) within the subpopulations defined in the first step.

We will discuss these two steps in detail in the next Sections.

Heterogeneity in the Conditional ITT

The BCF-IV algorithm starts from modifying (1.7) to adapt it for the estimation of the intention-to-treat, by including the instrumental variable Z_i:

$$Y_i = f(\mathbf{X}_i, Z_i)+\epsilon_i \approx \mathbb{T}_1(\mathbf{X}_i, Z_i)+...+\mathbb{T}_q(\mathbf{X}_i, Z_i)+\epsilon_i,\ \epsilon_i \sim \mathcal{N}(0,\sigma^2), \quad (1.8)$$

strong confounding. Moreover, since BCF includes the entire predictors' vector, $\mathbf{X}$, even if the propensity score is mis-specified or poorly estimated, the model allows for the possibility that the response remains correctly specified (P. R. Hahn et al., 2020). In Appendix 6, we show that even if the estimate $\hat{\pi}(x)$ of the propensity score is incorrectly specified the results are still widely robust.

where, for simplicity, we assume the error to be a mean zero additive noise as in P. R. Hahn et al. (2020), Hill (2011), Logan et al. (2019). The conditional expected value of Y_i can be expressed as:

$$\mathbb{E}[Y_i|Z_i = z, \mathbf{X}_i = x] = \mu(z, x), \tag{1.9}$$

and in turn the conditional intention-to-treat, $ITT_Y(x)$, is:

$$ITT_Y(x) = \mathbb{E}[Y_i|Z_i = 1, \mathbf{X}_i = x] - \mathbb{E}[Y_i|Z_i = 0, \mathbf{X}_i = x] = \mu(1, x) - \mu(0, x). \tag{1.10}$$

Then, adapting to an irregular assignment mechanism the model proposed by P. R. Hahn et al. (2020), we adopt the following functional form for (1.9):

$$\mathbb{E}[Y_i|Z_i = z, \mathbf{X}_i = x] = \mu(x, \pi(x)) + ITT_Y(x)z \tag{1.11}$$

where $\pi(x)$ is the propensity score for the instrumental variable:

$$\pi(x) = E[Z_i = 1|\mathbf{X}_i = x]. \tag{1.12}$$

The expression of $\mathbb{E}[Y_i|Z_i = z, \mathbf{X}_i = x]$ as a sum of two functions is central: the first component of the sum, $\mu(x, \pi(x))$, directly models the impact of the control variables on the conditional mean of the response (the component that is independent from the treatment effects) while the second component $ITT_Y(x)z$ models directly the intention-to-treat effect as a nonlinear function of the observed characteristics (this second components captures the heterogeneity in the intention-to-treat). Both the functions μ and ITT_Y are given independent priors. These priors are chosen in line with P. R. Hahn et al. (2020) to be for the first component the same priors of Chipman et al. (2010) (see Section 1.2). However, for the second component the priors are changed in a way that allows for less deep, hence simpler trees[11].

The expression of $\mathbb{E}[Y_i|Z_i = z, \mathbf{X}_i = x]$ as a sum of two functions has a double effect: (i) on the one hand, it allows the algorithm to learn which component in the heterogeneity of the conditional mean of the outcome is driven by a direct effect of the control variables and which component

[11] The depth penalty parameters are set to be $\eta = 3$ and $\beta = 0.25$ (instead of $\eta = 2$ and $\beta = 0.95$).

is the true heterogeneity in the effects of the assignment to the treatment Z_i on Y_i; (ii) on the other hand, it allows the predictions of the treatment effect driven by the BART to be modelled directly and separately with respect to the impact of the control variables (P. R. Hahn et al., 2020).

The estimated propensity score, in the BCF model, is not used for the estimation of the effects but is included, as an additional covariate, in the first component of (1.11) to mitigate possible problems connected to *regularization induced confounding* (RIC)[12] and *targeted selection*[13]. Moreover, in scenarios where the instrumental variable is not randomized ex-ante, the inclusion of the estimated propensity score, $\hat{\pi}(x)$, leads to an improvement in the discovery of the heterogeneity in the causal effect (P. R. Hahn, Dorie, et al., 2018). Furthermore, it is important to highlight that choosing a mis-specified definition of $\hat{\pi}(x)$ does not impact in a significant way the quality of the results as shown in Appendix 6. This is due to the fact that this first step of our algorithm is not about directly estimating the conditional CACE but is tailored to discover the heterogeneity in $ITT_Y(x)$.

Once one estimated with the BCF-IV the unit-level intention-to-treat, one can build a simple binary tree, using a CART model (Breiman, 1984), on the fitted values ($\widehat{ITT}_Y(\mathbf{X}_i)$) to discover the drivers of the heterogeneity. Such a two step approach (i.e.,first estimate the unit level effects, then discover the heterogeneity through a tree-based algorithm) is extremely valuable in providing interpretable information on the heterogeneity of the effects, and is investigated further in Chapter 2.

Estimation of Conditional CACE

Once the heterogeneous patterns in the intention-to-treat (ITT) are learned from the algorithm, one can estimate the conditional CACE,

[12] RIC is analyzed in depth in P. R. Hahn, Carvalho, et al. (2018). RIC issues rise when the ML algorithm used for regularizing the coefficient does not shrink to zero some coefficients due to a nonzero correlation between Z_i and X_i resulting in an additional degree of bias that is not under the researcher's control.

[13] Targeted selection refers to settings where the treatment (or in an IV scenario the assignment to the treatment) is assigned based on an ex-ante prediction of the outcome conditional on some characteristics $\mathbf{X}_i$. We refer to P. R. Hahn et al. (2020) for a discussion of targeted selection problems.

$\tau^{cace}(x)$. To do so, one can simply use the method of moments estimator in Equation (1.6) within all the different sub-populations that were detected in the previous step.
The conditional CACE can be estimated in a generic sub-sample (i.e., for each $\mathbf{X}_i \in \mathbb{X}_j$, where $\mathbb{X}_j$ is a generic node of the tree, like a non-terminal node or a leaf) as:

$$\hat{\tau}^{cace}(\mathbf{X}_i) = \frac{\widehat{ITT}_Y(\mathbf{X}_i)}{\hat{\pi}_C(\mathbf{X}_i)}, \tag{1.13}$$

where $\hat{\pi}_C(\mathbf{X}_i)$ is estimated as:

$$\hat{\pi}_C(\mathbf{X}_i) = \frac{1}{N_{1,l}} \sum_{l:X_l \in \mathbb{X}_j} W_l Z_l - \frac{1}{N_{0,l}} \sum_{l:X_l \in \mathbb{X}_j} W_l (1 - Z_l), \tag{1.14}$$

and $\widehat{ITT}_Y(\mathbf{X}_i)$ as:

$$\widehat{ITT}_Y(\mathbf{X}_i) = \frac{1}{N_{1,l}} \sum_{l:X_l \in \mathbb{X}_j} Y_l^{obs} Z_l - \frac{1}{N_{0,l}} \sum_{l:X_l \in \mathbb{X}_j} Y_l^{obs} (1 - Z_l), \tag{1.15}$$

where $N_{k,l}$ (where $k \in \{0, 1\}$) is the number of observations with $Z_l \in \{0, 1\}$ in the sub-sample of observations with $X_l \in \mathbb{X}_j$[14].

To show in detail how this second step works, let us use a toy example. Let's imagine a simple heterogeneity structure for $ITT_Y(x)$ where $ITT_Y(X_{i,p} > 0) \gg ITT_Y(X_{i,p} \leq 0)$ and $X_{i,p} \in (-1, 1)$ is a single regressor. This is namely, the case where the average intention-to-treat for those individuals with positive values of $X_{i,p}$ is greater than for individuals with non-positive values. Then, the conditional CACE can be estimated in the two different sub-populations defined with respect to X_p as[15]:

$$\hat{\tau}^{cace}(X_{i,p} > 0) = \frac{\widehat{ITT}_Y(X_{i,p} > 0)}{\hat{\pi}_C(X_{i,p} > 0)}, \tag{1.16}$$

$$\hat{\tau}^{cace}(X_{i,p} \leq 0) = \frac{\widehat{ITT}_Y(X_{i,p} \leq 0)}{\hat{\pi}_C(X_{i,p} \leq 0)}. \tag{1.17}$$

[14] It is worth highlighting that, since the supervised machine learning technique is used in the discovery phase and not in the estimation phase, the estimators that are proposed here could be used in a more "traditional way", in settings where the subgroups are defined ex-ante by the researcher.

[15] Alternatively, one can perform a Two Stage Least Squares (2SLS) regression within the different sub-populations. This is our preferred estimation strategy and is the one used both for the simulations and the real application.

1.4 Properties of the conditional CACE estimator

In the case of a binary instrument ($Z_i \in \{0,1\}$) and a binary treatment variable ($W_i \in \{0,1\}$), Angrist et al. (1996) and Imbens and Rubin (2015) revealed that the population versions of (1.13)-(1.15) correspond to a Two Stage Least Squares (henceforth referred to as 2SLS) estimator of τ^{cace}, in the cases where the four IV assumptions can be assumed to hold. Hence, since this case is analogous to our setting, one can apply the 2SLS method in every node $\mathbb{X}_j$ of the tree $\mathbb{T}$ for the estimation of the effect on the compliers population, as it is presented by Imbens and Rubin (1997).

The two simultaneous equations of the 2SLS estimator are, in the population,

$$Y_i^{obs} = \alpha + \tau^{cace} W_i + \epsilon_i, \tag{1.18}$$

$$W_i = \pi_0 + \pi_C Z_i + \eta_i, \tag{1.19}$$

where $\mathbb{E}(\epsilon_i) = \mathbb{E}(\eta_i) = 0$, and $\mathbb{E}(Z_i \eta_i) = 0$[16]. In the econometric terminology, the explanatory variable W_i is *endogenous*, while the IV variable Z_i is *exogenous*.

We can express the 2SLS equations, conditional on a subpopulation of a node $\mathbb{X}_j$, as

$$Y_{i,\mathbb{X}_j}^{obs} = \alpha_{\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} W_{i,\mathbb{X}_j} + \epsilon_{i,\mathbb{X}_j}, \tag{1.20}$$

$$W_{i,\mathbb{X}_j} = \pi_{0,\mathbb{X}_j} + \pi_{C,\mathbb{X}_j} Z_{i,\mathbb{X}_j} + \eta_{i,\mathbb{X}_j}, \tag{1.21}$$

where $\mathbb{E}(\epsilon_{i,\mathbb{X}_j}) = \mathbb{E}(\eta_{i,\mathbb{X}_j}) = 0$, and $\mathbb{E}(Z_{i,\mathbb{X}_j} \eta_{i,\mathbb{X}_j}) = 0$. Moreover, the following reduced equation (obtained plugging (1.21) into (1.20)) holds:

$$\begin{aligned} Y_{i,\mathbb{X}_j}^{obs} &= \left(\alpha_{\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} \pi_{0,\mathbb{X}_j}\right) + \left(\tau_{\mathbb{X}_j}^{cace} \pi_{C,\mathbb{X}_j}\right) Z_{i,\mathbb{X}_j} + \left(\epsilon_{i,\mathbb{X}_j} + \tau_{\mathbb{X}_j}^{cace} \eta_{i,\mathbb{X}_j}\right) \\ &= \bar{\alpha}_{\mathbb{X}_j} + \gamma_{\mathbb{X}_j} Z_{i,\mathbb{X}_j} + \psi_{i,\mathbb{X}_j}. \end{aligned} \tag{1.22}$$

In the case of a single instrument, the logic of IV regression is that one can estimate the respective parameters $\pi_{C,\mathbb{X}_j}$ and $\gamma_{\mathbb{X}_j} = \tau_{\mathbb{X}_j}^{cace} \pi_{C,\mathbb{X}_j}$ of the regressions (1.21) and (1.22) above by least squares, when the observations in each node are independent and identically distributed, then

[16] The latter comes from the fact that (1.19) is assumed to represent the linear projection of W_i onto Z_i.

obtaining an estimate of the parameter $\tau^{cace}_{\mathbb{X}_j}$ in (1.20). In particular, for every element $\mathbf{X}_i$ of a node $\mathbb{X}_j$, one can estimate $\tau^{CACE}(\mathbf{X}_i) = \tau^{cace}_{\mathbb{X}_j}$ through 2SLS, as the following ratio (Imbens & Rubin, 2015):

$$\hat{\tau}^{CACE}(\mathbf{X}_i) \equiv \hat{\tau}^{2SLS}_{\mathbb{X}_j} = \frac{\hat{\gamma}_{\mathbb{X}_j}}{\hat{\pi}_{C,\mathbb{X}_j}}. \tag{1.23}$$

The 2SLS estimator associated with (1.20)-(1.22) satisfies the next properties. They can be proved likewise in the application of 2SLS to the population case (see, e.g., Imbens and Rubin (2015)).

Theorem 1: Consistency of the Conditional 2SLS Estimator.
Let $\mathbb{E}(Z^2_{i,\mathbb{X}_j}) \neq 0$ (Assumption 1), $\mathbb{E}(Z_{i,\mathbb{X}_j}\epsilon_{i,\mathbb{X}_j}) = 0$ (Assumption 2) and $\pi_{C,\mathbb{X}_j} \neq 0$ (Assumption 3) hold. Then

$$\hat{\tau}^{2SLS}_{\mathbb{X}_j} - \tau_{\mathbb{X}_j} \xrightarrow{p} 0 \;\; as \;\; N_{\mathbb{X}_j} \to \infty, \tag{1.24}$$

where $\xrightarrow{p}$ denotes convergence in probability, and $N_{\mathbb{X}_j}$ is the number of observations within the node $\mathbb{X}_j$.

It should be noted that Assumption 3 is not necessarily guaranteed even if the overall instrument is strong. However, this assumption is standard in treatment effects variation papers such as the contribution of Ding et al. (2019)[17].

Theorem 2: Asymptotic Normality of the Conditional 2SLS Estimator.
Let Assumptions 1, 2, and 3 hold. Let also $\mathbb{E}(Z^2_{i,\mathbb{X}_j}\epsilon^2_{i,\mathbb{X}_j})$ be finite (Assumption 4). Then

$$\sqrt{N_{\mathbb{X}_j}}\big(\hat{\tau}^{2SLS}_{\mathbb{X}_j} - \tau_{\mathbb{X}_j}\big) \xrightarrow{d} \mathcal{N}\big(0, N_{\mathbb{X}_j} avar(\hat{\tau}^{2SLS}_{\mathbb{X}_j})\big) \;\; as \;\; N_{\mathbb{X}_j} \to \infty, \tag{1.25}$$

where $\xrightarrow{d}$ denotes convergence in distribution, $\mathcal{N}$ stands for normal distribution, and $avar(\hat{\tau}^{2SLS}_{\mathbb{X}_j})$ is the asymptotic variance of the 2SLS estimator. The

[17] In cases where Assumption 3 is not violated but the proportion of compliers within the nodes approaches zero, there could be potential problems related to heterogeneity driven by these small values of $\pi_{C,\mathbb{X}_j}$ and weak instruments issues within the nodes. In our scenario, since the heterogeneity in the treatment effect is detected with respect to the conditional ITT (which does not take directly into account $\pi_{C,\mathbb{X}_j}$ for its computation), our model is robust to possible treatment effects heterogeneity variations driven by smaller values of $\pi_{C,\mathbb{X}_j}$. Moreover, we do run weak-instrument tests within every node and we discard those nodes where a weak-instrument issue is detected.

asymptotic variance of the estimator is (Wooldridge, 2015):

$$avar(\hat{\tau}^{2SLS}_{\mathbb{X}_j}) = \frac{\hat{\psi}^2_{\mathbb{X}_j}}{TSS_{W\mathbb{X}_j}\rho^2_{W\mathbb{X}_j,Z\mathbb{X}_j}} \tag{1.26}$$

where $TSS_{W\mathbb{X}_j}$ is the conditional total sum of square of W, $\rho^2_{W\mathbb{X}_j,Z\mathbb{X}_j}$ is the conditional squared correlation of a regression of W on Z, and $\hat{\psi}^2_{\mathbb{X}_j}$ is estimated as:

$$\hat{\psi}^2_{\mathbb{X}_j} = \frac{1}{N_{\mathbb{X}_j} - 2}\sum_{i=1}^{N_{\mathbb{X}_j}} \hat{u}_i^2 \tag{1.27}$$

where u_i is the regression residual obtained regressing the outcome on W. The variance can be estimated using the finite-sample estimators of the corresponding components of equation 1.26.

The proofs of the two Theorems above directly follow from their unconditional versions[18]. In this case, for the convergence of our estimator to $\tau_{\mathbb{X}_j}$ and its normality to hold approximately we need to have a sufficient number of observations within every node. Hence, we suggest to perform our algorithm on sufficiently large datasets and to trim those nodes where the number of observations is not large enough[19].

2 Monte Carlo simulations

To evaluate the performance of the BCF-IV algorithm we compare it, using Monte Carlo Simulations, with two methods that are directly tailored for drawing causal inference in irregular assignment mechanism scenarios: the Honest Causal Trees with Instrumental Variable (HCT-IV) algorithm (Bargagli-Stoffi & Gnecco, 2020) and the Generalized Random Forests (GRF) algorithm (Athey et al., 2019). Both the latter algorithms outperform other machine learning methodologies which are not tailored for irregular assignment mechanisms (Bargagli-Stoffi & Gnecco,

[18] For further details on these proofs we refer to Wooldridge, 2002, Section 5.2.

[19] An R function for BCF-IV, and the code used for both the simulations and the application study in the Chapter are available at the following GitHub page: https://github.com/fbargaglistoffi.

2020), so we focus, in this context, just on a comparison within these three algorithms.

Since the foremost focus of this Chapter is on discovery of the heterogeneity in the causal effects, we compare the algorithms on three dimensions: (i) the correct choice of the variable that drives the heterogeneity (*heterogeneity driving variable* [HDV]), (ii) the correct choice of the threshold value used to perform the binary split of the data given the right identification of HDV, and (iii) the mean-squared error for the heterogeneous causal effects given the correct choice of HDV.

For Monte Carlo Simulations we build two different designs. The functional forms of the designs are built following the simulation designs in G. Wang et al. (2018). The first design takes the form of $Y_i = \sum_{p=1}^{k} X_{i,p} + W_i X_{i,1} + \xi_i + \epsilon_i$ where $X_{i,p} \sim \mathcal{N}(0,1)$, $W_i \sim Bern(0.5), \xi_i \sim \mathcal{N}(0, 0.01)$ and $\epsilon_i \sim \mathcal{N}(0,1)$. The interaction term between the regressor $X_{i,1}$ and the treatment indicator W_i is functional to heterogenise the treatment effects, while the nuisance parameter ξ_i is an unobserved variable that affects both W_i and the response variable Y_i. The second design has the same functional form but $x_{i,p} \sim Bern(0.5)$. In both the designs we set the correlations between W_i, the instrument Z_i and the nuisance parameter ξ_i to be: $Cor(W_i, Z_i) \in (0.55, 0.65)$ and $Cor(W_i, \xi_i) \in (0.45, 0.55)$, while k assumes values 5 and 10 and the sample sizes are 500, 1,000, 5,000. For both designs the results are aggregated over 30 rounds of simulations.

The results from the simulations are shown in Table 4. As shown in Panel A, the correct identification of the HDV is very similar for BCF-IV and GRF in the designs with 500 and 5,000 units. GRF is *faster* in identifying the right HDV, as it outperforms both BCF-IV and HCT-IV when the sample size is 1,000. Panel B depicts the results in terms of mean squared error between the true and the predicted threshold used to perform the binary split of the data. The threshold is not available in Design 2 where the regressors are binary variables. BCF-IV outperforms both GRF and HCT-IV with all the sample sizes and with both 5 and 10 features (with the exception of Design 1 in the sample of 5,000 units with 5 features). Panel C depicts the mean squared error of prediction for the causal effects given the correct identification of the HDV. Another clear advantage of using BCF-IV is given by the correct identification of the treatment effects. Indeed, BCF-IV outperforms, in terms of lower mean-squared-error of prediction for the treatment effects, the other algorithms in both the designs with 5 and 10 features (with the exception of Design 1 in the samples of 5,000 units with 5 features and 500 units with 10 features). Hence, in a scenario with binary regressors, BCF-IV is preferable irrespective of the sample size. However, in a scenario with continuous regressors[20], there is a trade-off between the capacity of getting the right HDV and the capacity of correctly estimating the causal effect. In designs with samples sizes of 1,000, GRF outperforms BCF-IV in correctly identifying the HDV but fails to precisely estimate the causal effect. Indeed, on one side GRF is directly built to detect the heterogeneity in the causal effects, but it depicts a lower precision in their estimation. On the other side, BCF-IV is not directly optimized to discover effect variation, but its two-stage procedure allows researchers to estimate the effects at a higher level of precision. We want to highlight that, as the sample size increases, particularly in the scenario with 5 features, both the algorithms get to the same large-sample results in terms of correct HDV identification and the mean-squared-error of prediction.

We argue that, in small samples, BCF-IV would be preferable to GRF because, while the proportion of correctly identified HDVs is very similar, the gains obtained both in terms of mean-squared-error between the true and predicted threshold and the true and predicted causal effects are much larger. In fact, the relative gap[21] between the true and predicted causal effects ranges between 15% and 81% in favour of BCF-IV, while the relative gap in the proportion of correctly identified HDV ranges from -10% (in favour of GRF) to 30% (in favour of BCF-IV). Hence, we claim

[20] For instance, when the regressors are distributed according to a standardized normal distribution.

[21] The formula for the relative gap is, for the MSE of prediction, the following (G. Wang et al., 2018):

$$\text{Relative Gap} = \frac{MSE_{GRF} - MSE_{BCF\text{-}IV}}{MSE_{GRF}} \times 100.$$

The relative gap is positive when BCF-IV outperforms GRF and negative vice versa.

that the gain in the mean-squared-error of prediction for the causal effect outweighs the slower identification of HDVs. This holds true as BCF-IV and GRF converge to a very similar fit, as the sample size increases, with respect to the three dimensions that are the object of our analysis. Moreover, the large sample behaviour of BCF-IV is slightly better than the one of all the other techniques.

Table 4: Monte Carlo comparison of BCF-IV, GRF and HCT-IV

Panel (A): proportion of correctly identified heterogeneity driving variables (HDV)

		Sample Size					
#Features	Approach	500	1,000	5,000	500	1,000	5,000
		Design 1			Design 2		
		HDV					
5	BCF-IV	0.57	0.57	**1.00**	0.90	0.96	**1.00**
	GRF	**0.63**	**0.83**	**1.00**	**0.93**	**1.00**	**1.00**
	HCT-IV	0.43	0.40	0.70	0.53	0.56	0.86
10	BCF-IV	**0.30**	0.33	0.73	**0.53**	0.93	**1.00**
	GRF	0.23	**0.43**	**1.00**	0.50	**0.96**	**1.00**
	HCT-IV	0.17	0.30	0.77	0.20	0.63	0.83

Panel (B): MSE between true and predicted threshold

		Sample Size					
#Features	Approach	500	1,000	5,000	500	1,000	5,000
		Design 1			Design 2		
		Threshold					
5	BCF-IV	**0.062**	**0.037**	0.006	-	-	-
	GRF	0.063	0.069	**0.002**	-	-	-
	HCT-IV	0.185	0.188	0.045	-	-	-
10	BCF-IV	**0.046**	**0.014**	**0.017**	-	-	-
	GRF	0.190	0.125	0.040	-	-	-
	HCT-IV	0.096	0.023	0.167	-	-	-

Panel (C): MSE between true and predicted causal effects

		Sample Size					
#Features	Approach	500	1,000	5,000	500	1,000	5,000
			Design 1			Design 2	
				Causal Effects			
5	BCF-IV	**0.047**	**0.026**	0.002	**0.048**	**0.005**	**0.010**
	GRF	0.330	0.030	**0.001**	0.055	0.006	0.011
	HCT-IV	0.067	0.051	0.012	0.048	0.005	0.010
10	BCF-IV	0.017	**0.021**	**0.020**	**0.036**	**0.005**	**0.002**
	GRF	0.230	0.197	0.128	0.190	0.105	0.012
	HCT-IV	**0.013**	0.039	0.046	0.036	0.005	0.002

Note: In these Tables we show the results from Monte Carlo simulations. Panel A depicts the proportion of correctly identified heterogeneity driving variables (HDVs). Panel B shows the mean squared error between true and predicted threshold (which is available just for the model with normally distributed variables). Panel C depicts the mean squared error between true and predicted causal effects. We highlighted in bold the best results for every round of simulations. In case of the same mean-squared-error of prediction (that occurs with BCF-IV and HCT-IV since both techniques are based on the same estimator) we assigned the best performance to the technique that performs better in identifying the correct HDVs. The results are obtained by aggregating 30 bootstrap samples.

In Appendix 6, we provide a number of robustness checks of the Monte Carlo simulation. In particular, we focus on what happens to the fit of the three algorithms when one: (i) changes the correlation between Z_i and W_i (possible weak-instrument problems)[22]; (ii) introduces a violation in the exclusion restriction; (iii) changes the specification of the propensity score for the BCF-IV; (iv) introduces multiple heterogeneity variables; (v) changes the error distribution. The results that we highlighted before hold true also in the robustness checks: BCF-IV converges slowly to an optimal identification of the HDVs but largely outperforms GRF with respect to the mean-squared-error of prediction for the causal

[22] It is important to highlight that in order to avoid weak-instrument problems within a node our algorithm performs a weak-instrument test in every sub-sample (namely, an F-test on the first stage regression) and discards the nodes where the null hypothesis of weak instrument is not rejected.

effects[23]. Moreover, the performance of BCF-IV does not seem to widely deteriorate, as compared to the baseline models in Table 4, in any of the robustness designs.

3 Heterogeneous causal effects of education funding

There is a wide consensus that education positively influences labor market outcomes (see the review by Psacharopoulos and Patrinos (2018)). Students' performance can be driven by multiple factors connected with students' characteristics and environmental characteristics. However, to the best of our knowledge, this is the first contribution to study the impact of additional school funding on students' performance using machine learning techniques tailored for causal inference. In this Section, we apply the BCF-IV algorithm to evaluate the impact and estimate the heterogeneity in the effects of additional funding to schools with disadvantaged students on students' performance. First, we describe the data used for this application. Next, we depict the identification strategy. Finally, we describe the results obtained and their relevance in the economics of education literature.

3.1 Data

Starting from the year 2002, the Flemish Ministry of Education promoted the "Equal Educational Opportunities" program (henceforth referred to as EEO) to ensure equal educational opportunities to all the students (OECD, 2017a). The EEO program provides additional funding for secondary schools with a significant share of disadvantaged students. Owing to the funding schools can hire additional teachers and increase the number of teaching hours. Pupils are considered to be disadvantaged on the basis of five different indicators: (i) the pupil lives outside the

[23] However, when we introduce a partial violation of the exclusion restriction assumption (design 2) we see exactly the opposite: BCF-IV outperforms GRF with respect to the identification of the correct HDV while GRF outperforms BCF-IV in precisely estimating the causal effects.

family; (ii) the pupil does not speak Dutch as a native language; (iii) the mother of the pupil does not have a secondary education degree; (iv) the pupil receives educational grant guaranteed for low income families; and (v) one of the parents is part of the travelling population. In order for a school to be eligible for the EEO funding, it needs to satisfy two conditions: the first condition is that the share of students with at least one of the five characteristics has to exceed an exogenously set threshold; to avoid fragmentation of resources, the second condition requires that the additional resources should be at least larger than six teaching hours a week. The exogenous cutoff is, for students in the first two years of secondary education (first stage students), a minimum share of 10% disadvantaged students.

The Flemish Ministry of Education provided us with data on the universe of pupils in the first stage of education in the school year 2010/2011 (135,682 students). In particular, we have data on student level characteristics and school level characteristics. The student level characteristics cover the gender of the pupil (*gender*), the grade retention in primary school (*retention*) and the inclusion of the pupil in the special needs student population in primary school (which serves as a proxy of student's low cognitive skills). The school level characteristics include both the teacher characteristics, such as the teachers' age, seniority and education, in addition to principal characteristics, such as the principals' age and seniority. Teacher and principal seniority measure the level of experience of the teachers and principals, respectively. These variables assume values in the range of 1 to 7, where the teachers (and principals) with a seniority level of 1 are the least experienced (0-5 years of experience) and teachers (and principals) with a seniority level of 7 are the most experienced (more than 30 years of experience)[24]. Similarly, the ages of teacher and principal are reported as categorical variables that range from 1 to 8, where teachers/principals in the first category are the youngest (less than 30 years old) and teachers/principals in the last category are the

[24] Teachers and principals' seniority classes are the following: class 1: between 0 and 5 years of experience; class 2: between 6 and 10; class 3: between 11 and 15; class 4: between 16 and 20; class 5: between 20 and 25; class 6: between 26 and 30; class 7: more than 30.

oldest (more than 60 years old)[25]. Teachers' education records whether or not the teacher holds a pedagogical training (in the following we will refer to it as "teacher training"). All these variables are aggregated at school level in the form of averages (for age and seniority) and shares (for teachers' education) and assigned to each student with respect to the school where he/she is enrolled.

The outcome variables are two dummy variables defined as follows: the variable *progress school* assumes value 1 if the student progresses to the following year without any grade retention and 0 if not (this variable is a complement of school retention); the variable *A-certificate* assumes value 1 if the student gets an "A-certificate" at the end of the school year (which is the most favorable outcome) and 0 if not. Since we do not have data on standardized test scores for Flemish students, *A-certificate* is a good, available proxy of student performance. Every year, each student performs a final test and gets a ranking from "A" to "C". Students that get an "A" can progress school without any restriction, while the students that get either "B" or "C" can progress school but only in specific programs or have some grade retention. Both these outcome variables are proxies for different levels of students' performance: a positive *A-certificate* proxies for a higher level of performance than a positive *progress school*. In principle, the target of a policy-maker could be to have the highest possible share of students getting "A-certificates" and the lowest share of students not progressing through school.

3.2 Identification strategy

To evaluate the impact of the policy on students' performance we apply the BCF-IV within a regression discontinuity design (J. Hahn et al., 2001). Regression Discontinuity Design (RDD) is a method that aims at evaluating the causal effects of interventions in settings where the assignment to the treatment is determined (at least partly) by the values of an observed

[25] Teachers and principals' age classes are the following: class 1: less than 30 years old; class 2: between 30 and 34; class 3: between 35 and 39; class 4: between 40 and 44; class 5: between 45 and 49; class 6: between 50 and 54; class 7: between 55 and 60; class 8: more than 60.

covariate lying on either sides of a threshold point. The idea is that subjects just above and below this threshold are very similar and one can assume a quasi-randomization around the threshold (Mealli & Rampichini, 2012). RDDs are categorized in sharp RDDs and fuzzy RDDs. In sharp RDDs, the central assumption is that, around the threshold, there is a sharp discontinuity (from 0 to 1) in the probability of being treated. This is due to the fact that in sharp RDDs there is no room for imperfect compliance. In many real world scenarios, however, thresholds are not strictly implemented, as in the case of our application. To deal with these situations, one can use fuzzy RDDs, which are applicable when around the threshold the probability of being actually treated changes discontinuously, but not sharply from 0 to 1 (i.e., the jump in the probability of being treated is less than 1). In our application of the fuzzy RDD technique, we exploit two cutoffs around the 10% share of disadvantaged students in the first stage of secondary education. The students in schools just below the threshold are assigned to the control group ($Z_i = 0$), while the students in schools just above the threshold are assigned to the treated group ($Z_i = 1$). The bandwidth around the threshold (from which one obtains the two cutoffs) is determined using the *"rdrobust package"* in R (Calonico et al., 2015). The optimal, bias-corrected bandwidths around the threshold are 3.5% and 3.7%, respectively for the outcome variables *A-certificate* and *progress school*. Accordingly to these two bandwidths, we obtain two refined samples where the sample with the 3.5% bandwidth is the smallest and the sample with the 3.7% bandwidth the largest. We run a series of robustness checks for the selection of the bandwidths around the cutoff. In order to validate the bandwidths selected using the method of Calonico et al. (2015), we run additional analyses implementing the Bayesian methods proposed by Li et al. (2015) and by Mattei and Mealli (2016). In these papers, the authors implement a hierarchical Bayesian model for assessing the balance of the covariates between the groups of observations assigned to the treatment and the ones assigned to the control. For both the thresholds selected following Calonico et al. (2015) (i.e., 3.5% and 3.7%), the probability of the pre-assignment variables being well-balanced is high for the subpopulations defined by values of the cutoff strictly lower than 3.5% and 3.7%. Indeed, these probabilities are larger than or close to 0.8, indicating that the covariates are balanced in the two groups.

Moreover, to guarantee an equal representation to all the schools, and to avoid biases related to the over-representation of biggest schools' students, we sample 50 pupils from each school. In turn, this leads to a higher balance among the averages between the observations assigned to the treatment and the observations assigned to the control, as shown in panel (a) of Figure 5. In Appendix 7, we run a series of tests to show that the RDD (Regression Discontinuity Design) is valid for this application. Moreover, as a robustness check we sample a higher number of students according to the size of the smallest school (62 pupils) from every school[26]. In Appendix 8, we show the balance in the samples of units assigned to the treatment and to the control in the second scenario.

3.3 Results

This Section assesses the effects of the additional funding on students' performances and highlights the main drivers of the heterogeneity in causal effects. These analyses are made for both the outcome variables: *A-certificate* and *progress school*[27].

A-Certificate

Proceeding from the seminal contributions of Coleman (1966) and Hanushek (2003) to recent contributions by Jackson et al. (2015) and Jackson (2018), the question on whether or not school spending affects students' performances has been central in the economic literature.

In our study, the variable *A-certificate* serves as a proxy for positive performance. In our sample, the students that got an "A-certificate"

[26] We do not apply any sampling with replacement technique to avoid oversampling of observations from smaller schools. Indeed, such a procedure may induce biases in the statistical *representatives* of the sample used to draw causal inference.

[27] It is important to highlight that the results for both the outcomes, considered separately, in terms of effects and heterogeneity drivers, remain roughly the same when we widen the sample of units included in the analysis (results are reported in the Appendix 8).

are the 91.73% of the total population. In Figure 3, the heterogeneous Complier Average Causal Effects (CACE) estimated using the proposed model are depicted[28] [29]. The darker the shade of blue in the node the higher the causal effect.

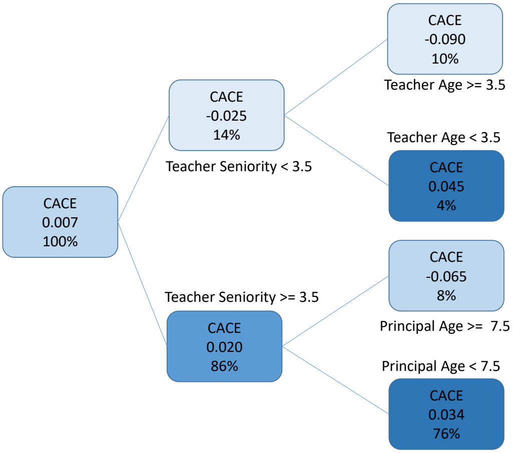


Figure 3: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *A-certificate* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in (P. R. Hahn et al., 2020). The overall sample size is 4,300. The significance level is * for a significance level of 0.1, ** for a significance level of 0.05 and *** for a significance level of 0.01.

Although positive, the overall effect of the additional funding is not significant. This finding is in line with the recent literature on school spending and students' performance in a cross-country scenario (Hanushek et al., 2016; Hanushek & Woessmann, 2017) and in the Flanders, in particular (De Witte et al., 2018). Nevertheless, rather than fo-

[28] The nodes for whom (i) it was not possible to compute the CACE or, (ii) the weak-IV test was not rejected were excluded from the plot.

[29] In Figures 3, 4, 8, 9, the so-called summarizing trees (P. R. Hahn et al., 2020) are depicted. A summarizing tree is a classification or regression tree that is built using the fitted values estimated from the BCF-IV. These summarizing trees are used to provide a visualization of the heterogeneity in the causal effects.

cusing on the overall average effect it is more interesting to explore the heterogeneous effects.

The first driver in the heterogeneity of the effects is the variable *teacher seniority*: for students in schools with more senior teachers, the effects of funding are larger. These results, even if not significant, show that the treatment effects are higher for students that are in schools with teachers that have more than 16 years of experience. The second driver of heterogeneity is the age of the teacher: students in schools with younger teachers (namely, when *teaching age* assumes values lower or equal to 3.5 on a scale from 1 to 8, referring to teachers younger than 40 years old) have an increase in their performance even if they are in schools with less experienced teachers. Both these heterogeneity drivers, namely, the seniority and age of the teacher, are particularly appreciable, as there are evidences in the education literature that connect teachers' seniority (Harris & Sass, 2011; Rice, 2010) and teachers' age (Holmlund & Sund, 2008) to their teaching performance, and in turn teaching performance to students' positive achievements (Goldhaber & Hansen, 2010).

Further heterogeneous effects come from the interaction between teacher's seniority and principal's age. The effect for students in schools with more experienced teachers and principals younger than 60 is higher than the effect on students in schools with teachers with similar experience but older principals. This evidence can be interpreted in the following way: the additional funding has a positive, but not statistically significant, effect in boosting the performance of students in the overall population, but it increases its effect in a notable way for those students in schools with more senior teachers and younger teachers and principals. These results are in line with the evidence that additional school funding does not boost the performance of the overall population of students (De Witte et al., 2018; Hanushek et al., 2016; Hanushek & Woessmann, 2017) and with the literature that connects students' achievements with teaching performance (Harris & Sass, 2011; Holmlund & Sund, 2008; Rice, 2010) and, in turn, teaching performance with students' performance (Goldhaber & Hansen, 2010). It is important to highlight that even if we find some evidence of treatment effects variation connected to teachers' seniority and age and principal age, the conditional causal effects are not significant.

Progress in School

The second outcome variable, *progress school*, assumes value 1 if the student progresses to the following year without any grade retention and 0 if not: roughly 98% of the students in the sample manage to progress in school in the first two years of secondary education. For the students unable to progress in school, this variable is used as a proxy of negative achievements. Therefore, it is relevant to understand if additional funding was effective in driving students away from negative performance. Figure 4 depicts the heterogeneous conditional CACEs: the darker the shade of green in the node, the higher the causal effect.

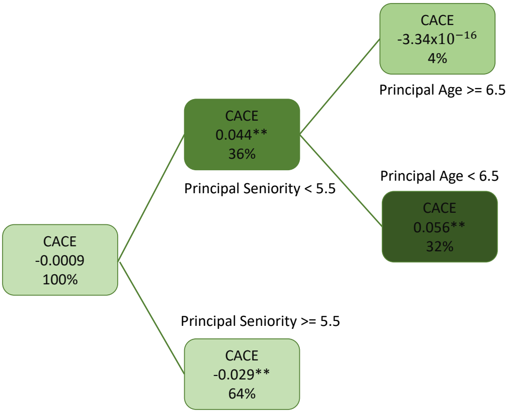


Figure 4: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *Progress School* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in (P. R. Hahn et al., 2020). The overall sample size is 4,450. The significance level is * for a significance level of 0.1, ** for a significance level of 0.05 and *** for a significance level of 0.01.

The additional funding has a slightly negative, but statistically insignificant, impact on the chance of progress in school for the overall students in the sample (again, this is in line with what was found by De Witte et al. (2018) at the school level). However, it is compelling to observe the main drivers of the heterogeneity in the causal effect. The first driver is the seniority level of principals: the treatment effect is positive and statistically significant (the effect in this case is 0.044**, meaning that being treated leads to an increase of 4.4% in the probability of progressing through school) for students in schools with less senior principals (less than 25 years of experience), and it is negative and statistically significant for students in schools with more experienced principals (the conditional effect is -0.029**, meaning that being treated leads to a decrease of 2.9% in the probability of progressing through school).

The second driver of the heterogeneity in CACE is the principal age. As in the case of the previous outcome, students in schools with younger principals (namely, principals younger than 55 years) show higher causal effects (the conditional effect is 0.056**). This holds true even when we do not condition on the principal's seniority (in this case the conditional effect is 0.062*). The sub-populations of students with positive and statistically significant effect account for the 36% and 32% of the overall sample (respectively, when conditioning, or not, on the *principal seniority* variable) and show effects that are in their absolute values between 49 and 62 times larger than the overall effect[30].

This evidence of higher and statistically significant effects of the funding in schools for those schools with younger and less senior principals (with respect to the average seniority and age, respectively) is a novel finding of this research. There is a compelling evidence in the literature regarding the role of principals in driving higher students' achievements (Eberts & Stone, 1988; Gentilucci & Muto, 2007), however this is, to the

[30] 49 is the ratio between the conditional treatment effect for the sub-population of students in schools with principals with less than 25 years of experience (0.0440) and the absolute value of the average treatment effect for the overall population (0.0009). 62 is the ratio between the conditional treatment effect for the sub-population of students that are in schools with less senior and younger principals (0.0557), corresponding to the darker shade of green leaf in Figure 4, and the absolute value of the average treatment effect on the overall population (0.0009).

extent of our knowledge, the first research that highlights the role of principal's age and seniority as drivers of treatment effect variations. Clearly, these characteristics could possibly correlate with unobservables, such as the effectiveness of principals (which may decrease as principals grow older). In any case, this finding opens up new fields for further investigation, in line with the newly established role of machine learning in the economic literature as a "theory-driving/theory-testing" tool (Mullainathan & Spiess, 2017).

4 Discussion

This Chapter developed a novel Bayesian machine learning technique, BCF-IV, to draw causal inference in scenarios with imperfect compliance. By investigating the heterogeneity in the causal effects, the technique expedites targeted policies. We manifested that the BCF-IV technique outperforms other machine learning techniques tailored for causal inference in precisely estimating the causal effects and converges to an optimal asymptotic performance in identifying the heterogeneity driving variables (HDVs). Moreover, using Monte-Carlo simulations, we showed that the competitive advantages of using BCF-IV, as compared to GRF or HCT-IV, are substantial. Peculiarly, the performance of BCF-IV in precisely estimating the heterogeneous causal effects shadows its slower convergence to an optimal identification of HDVs as compared to GRF. This is especially true if we look at the relative gaps between the BCF-IV and the other techniques.

BCF-IV can assist the researchers to shed light on the heterogeneity of causal effects in IV scenarios in order to provide to policy-makers a relevant knowledge for targeted policies. In our application, we evaluated the effects of additional funding on students' performances. While the overall effects are positive but not significant, there are significant differences among different sub-populations of students. Indeed, for students in schools with less senior and younger principals (with respect to the average seniority and age, respectively) the effects of the policy are between 49 and 62 times bigger than the effects on the overall population (in the most conservative scenario), and significant for the *Progress School* output.

On one hand, as an underlying mechanism, the need for additional funds can be higher in schools with younger teachers and principals, who are more often observed in the most disadvantaged schools. This phenomenon arises as senior teachers and principals select themselves out of the most disadvantaged schools and more into advantaged schools, thereby creating relatively more vacancies in disadvantaged schools. Therefore, on average, younger teachers and principal lack a real choice but to start working in the most disadvantaged schools. Moreover, owing to the additional funds, schools could use the funds to reduce class sizes, which might be more effective for younger (and less senior) teachers. On the other hand, we can think of the motivation for both teachers and principals to decrease as they grow older and this, in turn, have an impact on their performance. Favouring this hypothesis Ololube (2006) finds that motivation enhances productivity and has an impact on teachers' performances. However, teachers' motivation might positively and significantly affect teacher's job satisfaction, but it might not affect their performance. To the best of our knowledge, this is the first study that investigates the effects of age and seniority of principals on enhancing the effectiveness of school funding on students' performance. The investigation of the true causal channel is beyond the goals of this Chapter and is left to further investigation where more granular teachers' and principals' characteristics are available.

These results are relevant to the policy as they furnish the instruments to policy-makers to enhance the effects of additional funding on students' performance. Indeed, on one side policy-makers could target just students in school with positive, statistically significant effects reducing the overall costs of the policy and using the savings to experiment more effective policies in the other schools. On the other side, policy-makers could analyze the reason of lack of the effectiveness of funding in schools with certain characteristics and implement policies to boost the effects of future funding. Furthermore, the added value of our algorithm is that it could enable policy-makers to target just those units that benefit the most from the treatment and, at the same time, it provides an insight on possible inefficiencies in the allocation and/or usage of funding. From our analysis it seems that there is room for policies that support less senior principals since students in their schools show higher returns in terms of performance from additional funding.

Moreover, it would be of interest to extend the proposed analyses from the first to the second stage of secondary education. In particular, these analyses would provide further insights for at least two reasons: (i) the compliance status may change driving to different local effects since the exogenously set threshold to get funded is different, (ii) the different age of students could have an impact on the discovered heterogeneous effects.

The extension of these methods to other fields of economic investigation and the development of novel machine learning algorithms for targeted policies and welfare maximization can form the future scope of further research. In particular, the development of an algorithm that could deal with welfare maximization in the context of multiple outcomes is of interest. The "usual" Bayesian way for estimating CACE is via a data augmentation scheme (e.g., imputing compliance status and estimating impacts among estimated compliers [Imbens and Rubin, 2015]). In our algorithm we do not implement such a methodology, however it could be extended including a data augmentation scheme. Moreover, further research should focus on connecting BCF-IV and GRF into a single ensemble algorithm, following Grimmer et al. (2017), Van der Laan et al. (2007), to obtain a novel algorithm that combines the small and large sample properties of both BCF-IV and GRF to obtain possible gains in imperfect compliance scenarios.

Supplementary material for Chapter 1

5 Discussion on the IV assumptions

In a typical IV scenario one can express the treatment received as a function of the treatment assigned: $W_i(Z_i)$. This leads to distinguish four sub-populations of units (G_i) (Angrist et al., 1996; Imbens & Rubin, 2015): (i) those that comply with the assignment (*compliers*: $G_i = C : W_i(Z_i = 0) = 0$ and $W_i(Z_i = 1) = 1$); (ii) those that never comply with the assignment (*defiers*: $G_i = D : W_i(Z_i = 0) = 1$ and $W_i(Z_i = 1) = 0$); (iii) those that even if not assigned to the treatment always take it (*always-takers*: $G_i = AT : W_i(Z_i = 0) = 1, W_i(Z_i = 1) = 1$); (iv) those that even if assigned to the treatment never take it (*never-takers*: $G_i = NT : W_i(Z_i = 0) = 0, W_i(Z_i = 1) = 0$). In such a scenario what "one directly gets from the data" is the so-called Intention-To-Treat (ITT_Y):

$$ITT_Y = \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0], \tag{1.28}$$

which is defined as the effect of the intention to treat a unit on the outcome of the same unit. (1.28) can be written as the weighted average of the intention-to-treat effects across the four sub-populations of compliers, defiers, always-takers and never-takers:

$$ITT_Y = \pi_C ITT_{Y,C} + \pi_D ITT_{Y,D} + \pi_{NT} ITT_{Y,NT} + \pi_{AT} ITT_{Y,AT}, \tag{1.29}$$

where $ITT_{Y,G}$ is the effect of the treatment assignment on units of type G and π_G is the proportion of units of type G.

ITT_Y does not represent the effect of the treatment itself but just the effect of the assignment to the treatment. If we want to draw proper causal inference in such a scenario we need to invoke the four classical IV assumptions (Angrist et al., 1996):

1. *exclusion restriction*: $Y_i(0) = Y_i(1), \ \text{for } G_i \in \{AT, NT\}$ where, for each sub-population and $z \in \{0,1\}$, the shortened notation $Y_i(z)$ is used to denote $Y_i(z, W_i(z))$

2. *monotonicity*: $W_i(1) \geq W_i(0) \rightarrow \pi_D = 0$;

3. *existence of compliers*: $P(W_i(0) < W_i(1)) > 0 \rightarrow \pi_C \neq 0$;

4. *unconfoundedness of the instrument*:
$Z_i \perp\!\!\!\perp (Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), W_i(0), W_i(1))$.

In our application, these four assumptions are assumed to hold. Let us look at them in detail. The exclusion restriction is assumed to hold since we can reasonably rule out a direct effect of being eligible (around the threshold) on the performance of students. The effect, in this case, can be reasonably assumed to go through the actual reception of additional funding. Monotonicity holds by design: since we are in a one-sided non-compliance scenario there is no possibility for those who are not assigned to the treatment to defy and get the treatment. The same can be said about the existence of compliers. Since the sub-populations of always-takers and defiers can be ruled out by design, this leads to the fact that units receiving the treatment are compliers. Unconfoundedness of the instrument can also reasonably be assumed to hold since observations around the exogenous threshold are as good as if they were randomized to the assigned-to-the-treatment group and the assigned-to-the control group. This holds true especially since we do not observe any manipulation around the threshold and sorting of the units into the treated group.

6 Robustness checks in Monte Carlo simulations

We introduce some changes in the synthetic models used to test the fit of the BCF-IV (as compared to GRF and HCT-IV). The model from which we start is the simplest model introduced in Section 2: $Y_i = \sum_{p=1}^{k} X_{i,p} + W_i X_{i,1} + \xi_i + \epsilon_i$ where $X_{i,p} \sim \mathcal{N}(0,1), W_i \sim Bern(0.5), \xi_i \sim \mathcal{N}(0,0.01)$, $\epsilon_i \sim \mathcal{N}(0,1)$ and $k = 5$. We introduce 5 different variations in this model (each one corresponds to a different design in Table 9):

1. we change the correlation between Z_i and W_i in order to introduce possible weak-instrument problems: we decrease the correlation to $Cor(W_i, Z_i) \in (0.45, 0.55)$ and we do so by introducing in half of the population a very weak instrument $Cor(W_i, Z_i | X_{i,5} < 0) \in (0.35, 0.45)$;

2. we introduce a partial violation in the exclusion restriction;

3. we introduce multiple heterogeneity driving variables (HDVs):

$$Y_i = \sum_{p=1}^{k} X_{i,p} + \sum_{p=1}^{2} (W_i X_{i,p}) + \xi_i + \epsilon_i, \tag{1.30}$$

   where this variation is introduced to test if the HDVs are correctly selected even when they are multiple;

4. we change the error distribution, $\epsilon_i \sim \mathcal{U}(0,1)$, to test if the algorithm is robust to changes in the noise parameter;

5. we manipulate the propensity score function for the BCF-IV, to test if this model is robust to a mis-specification of $\hat{\pi}(x)$.

Table 5: Robustness checks

		Sample Size								
#Design	Approach	500	1000	5000	500	1000	5000	500	1000	5000
			HDV			Threshold		MSE given HDV		
1	BCF-IV	0.37	0.63	0.83	**0.065**	**0.031**	**0.007**	**0.078**	**0.117**	0.017
	GRF	**0.56**	**0.73**	**1.00**	0.157	0.107	**0.007**	0.230	0.122	**0.011**
	HTC-IV	0.23	0.36	0.73	0.289	0.090	0.065	0.110	0.140	0.022
2	BCF-IV	**0.63**	0.40	**0.77**	**0.022**	**0.061**	**0.002**	**0.005**	0.107	0.043
	GRF	0.43	**0.57**	0.23	0.052	0.067	0.003	0.171	**0.082**	**0.035**
	HTC-IV	0.43	0.37	0.53	0.189	0.170	0.064	0.018	0.102	0.056
3	BCF-IV	0.60	0.76	**1.00**	0.352	0.242	0.021	**0.183**	**0.077**	**0.004**
	GRF	**0.77**	**1.00**	**1.00**	**0.169**	**0.230**	**0.004**	0.776	0.365	0.275
	HTC-IV	0.53	0.63	0.73	0.323	0.289	0.180	0.312	0.116	0.031
4	BCF-IV	0.63	0.80	**1.00**	**0.043**	0.065	0.002	**0.006**	**0.009**	0.002
	GRF	**0.80**	**0.97**	**1.00**	0.068	**0.014**	**0.001**	0.211	0.031	**0.001**
	HTC-IV	0.47	0.40	0.70	0.207	0.103	0.087	0.029	0.014	0.017
5	BCF-IV	0.53	0.63	**1.00**	0.112	**0.020**	0.016	**0.018**	**0.011**	0.007
	GRF	**0.63**	**0.83**	**1.00**	**0.062**	0.069	**0.002**	0.330	0.030	**0.001**
	HTC-IV	0.43	0.40	0.70	0.185	0.188	0.045	0.067	0.051	0.012

Note: HDV refers to the proportion of correctly identified Heterogeneity Driving Variables (HDV); Threshold refers to the mean-squared-error between the true threshold and the predicted one; MSE given HDV refers to the mean-squared-error of prediction for the true causal effects. We highlighted in bold the best results for every round of simulations.

The results from these five different designs are reported in Table 5. In the presence of a weak-instrument (design 1), the fit of all the three algorithms deteriorates. As we saw in the Monte Carlo simulations in Section 2, GRF is better in identifying the correct HDV but BCF-IV outperforms both GRF and HCT-IV in picking the right threshold and in precisely estimating the causal effect. As we introduce a partial violation of the exclusion restriction (design 2), BCF-IV outperforms the other algorithms with respect to all the dimensions both in the cases with small sample and large sample sizes. When we introduce multiple heterogeneity driving variables, the capacity of correctly estimating the causal effects for GRF deteriorates while BCF-IV outperforms the other algorithms. In the last two designs (design 4 and 5), we again see a trade-off, for the designs with 500 and 1,000 units, between the capacity of correctly identifying the HDV (GRF outperforms the other techniques) and precisely estimating the causal effects (BCF-IV outperforms the other algorithms). In both the latter designs, BCF-IV and GRF get to fairly similar asymptotic results.

7 RDD Checks

In order to check whether or not the RDD (Regression Discontinuity Design) setting is valid, we implement the following checks (D. S. Lee & Lemieux, 2010)[31]: (i) we check the balance in the sample of units assigned to the treatment just above and below the threshold (this is done to check if the randomization holds); (ii) we examine if there are manipulations in the distribution of schools with respect to the share of disadvantaged students around the threshold, (iii) we employ a formal manipulation test, the McCrary test (McCrary, 2008), to discover potential sorting around the threshold; (iv) we check if there is a discontinuity in the probability of being assigned to the treatment around the threshold. Table 6 shows that the averages of the control variables are not statistically different for the group of units assigned to the treatment and assigned to the control around the cutoff, with the exception of *teacher seniority*. Thus, there is evidence that more senior teachers self-select in schools with lower disadvantaged students. However, as we will show in Section 3.3, this variable does not surface as a driver of significant heterogeneity in the estimated causal effects. We argue that including this variable in our model results in more robust findings. This is due to the fact that our model is robust to *spurious* heterogeneity coming from unbalances in the samples, as shown by P. R. Hahn et al., 2020 in randomized and regular assignment mechanisms' scenarios. Moreover, panel (b) of Figure 5 shows the standardized difference in the means for these two groups with the relative standardized confidence intervals. The McCrary manipulation test implemented in Calonico et al., 2015 through a Local-Polynomial Density Estimation leads to the rejection of the null hypothesis of the threshold manipulation[32]. Both these results and the plot of the distribution of schools with respect to the share of disadvantaged students around the threshold in

[31] The checks depicted in this Subsection are made on the sample of 50 students introduced in Subsection 3.2.

[32] The McCrary test leads to a T-value of -0.7497 corresponding to a p-value of 0.4534. The test is performed aggregating the student data at school level.

Figure 6 indicate that there is no evidence of manipulation. Finally, Figure 7 shows a clear discontinuity in the probability of being assigned to the treatment around the threshold.

However, as we pointed out in Section 3, schools that are assigned to the treatment actually *receive* the treatment if they satisfy an additional condition of a minimum of six teaching hours. This leads to a fuzzy-regression discontinuity design where the jump in the probability of being assigned to the treatment around the cutoff is not sharp. This scenario is characterized by imperfect compliance.

Students can be sorted, with respect to their compliance status, into two types: (i) students in schools above the cutoff with more than six teaching hours or students in schools below the cutoff (*compliers*: $W_i(Z_i = 1) = 1$ or $W_i(Z_i = 0) = 0$); (ii) students in schools above the cutoff but with less than six teaching hours (*never-takers*: $W_i(Z_i = 1) = 0)$[33].

The assignment to the treatment variable (i.e., studying in a school just below or above the cutoff) is a relevant instrumental variable in our scenario (namely, the correlation between Z_i and W_i is roughly 0.62). Moreover, we can reasonably assume both the exogeneity condition and the exclusion restriction to hold in this situation. On one side, since the randomization of the instrument holds there is no reason not to assume conditional independence between the instrument and the unobservables. On the other side, the exclusion restriction seems to hold as well since we can believe that being just below or above the threshold does not affect the performance of students in any way other than through the additional funding. In this imperfect compliance setting, the causal effect of the additional funding on the students' performance can be assessed through the Complier Average Causal Effect in (1.5). Moreover, using our novel BCF-IV algorithm we can estimate the Conditional Complier Average Causal Effect, (1.6), to assess the heterogeneity in the causal effects.

[33] This a so-called case of one-sided-non-compliance, in which we do not observe any *always-takers* since for those that are sorted out of the assignment to the treatment ($Z_i = 0$) there is no possibility to access the treatment.

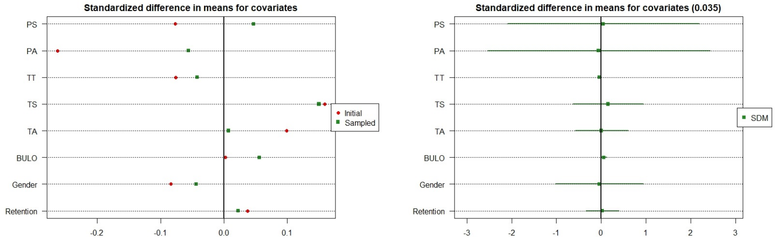


Figure 5: Balance improvement. The label "PS" refers to Principal Seniority, the label "PA" to Principal Age, the label "TS" to Teacher Seniority, the label "TA" to Teacher Age, the label "TT" to Teacher Training and the label "BULO" refers to students with special needs in primary education. (Left) Balance improvement obtained with sampling. "Initial" refers to the initial sample, while "Sampled" refers to the bootstrapped sample. (Right) Standardized difference in means (SDM) and 95% confidence interval around the cutoff with a bandwidth of 3.5%.

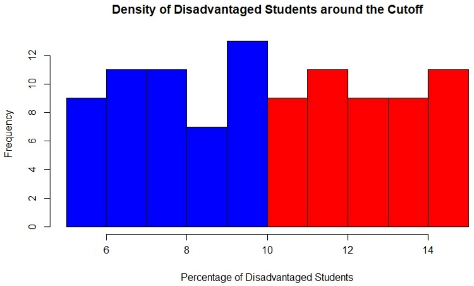


Figure 6: Frequency distribution of disadvantaged students around the threshold (10%). In red the density of the disadvantaged students in the units assigned to the treatment and in blue the density for the units assigned to the control. The densities are aggregated at school level.

	Above Cutoff		Below Cutoff		Full Sample		p-value
Retention	0.036	(0.187)	0.037	(0.189)	0.037	(0.188)	0.913
Gender	0.492	(0.500)	0.471	(0.499)	0.482	(0.500)	0.155
Special Needs	0.000	(0.000)	0.002	(0.044)	0.001	(0.030)	0.045
Teacher Age	4.022	(0.333)	4.024	(0.269)	4.023	(0.304)	0.814
Teacher Seniority	3.867	(0.452)	3.927	(0.342)	3.895	(0.404)	0.000
Teacher Training	0.982	(0.025)	0.981	(0.026)	0.982	(0.026)	0.169
Principal Age	6.022	(1.308)	5.951	(1.229)	5.988	(1.271)	0.067
Principal Seniority	5.778	(1.228)	5.829	(0.935)	5.802	(1.098)	0.120
Observations	2250		2050		4300		

Table 6: Results for 3.5% discontinuity sample with bootstrapped samples of size 50. Standard deviations are in parentheses and the p-value corresponds to a t-test for the difference between the means in the group above and below the threshold.

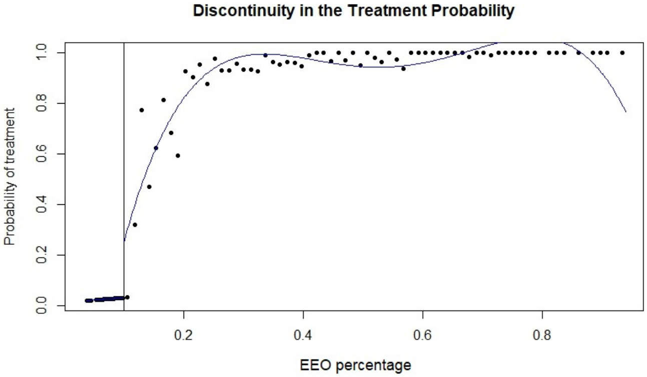


Figure 7: Probability of treatment given the share of disadvantaged students (EEO percentage) in the first stage of secondary education (cutoff 10%).

8 Robustness checks for policy evaluation

This Appendix tests the robustness of our models to sampling variations. The sampling variations introduced come from the following two sources: (i) a wider bandwidth around the threshold (changing from 3.5% to 3.7%); (ii) an expansion in the number of sampled units (from

50 up to the lowest number of students per school, which is 62). Moreover, an algorithm which detects the heterogeneity in the ITT effects is applied. To understand if the balance and the results are robust, we manifest the balance in the averages in the samples of units assigned to the treatment and assigned to the control (Tables 7, 8, 9) and the results of the causal effects when we increase the number of units sampled (Figures 8).

In all the different samples the school level characteristics remain widely balanced (with the exception of teacher seniority[34]). *Primary retention* and *Gender* seem to be slightly unbalanced when we widen the bandwidth, this however holds true just in the case where we sample through bootstrap 50 units (*Gender* in this case gets back to a good balance).

	Above Cutoff		Below Cutoff		Full Sample		p-value
Gender	0.471	(0.499)	0.493	(0.500)	0.482	(0.499)	0.110
Retention	0.039	(0.194)	0.035	(0.184)	0.037	(0.189)	0.418
Special Needs	0.001	(0.039)	0.000	(0.000)	0.001	(0.027)	0.045
Teacher Age	4.024	(0.269)	4.002	(0.333)	4.023	(0.304)	0.793
Teacher Seniority	3.926	(0.341)	3.867	(0.452)	3.895	(0.404)	0.000
Teacher Training	0.982	(0.025)	0.981	(0.026)	0.982	(0.026)	0.126
Principal Age	5.951	(1.228)	6.002	(1.308)	5.988	(1.271)	0.041
Principal Seniority	5.829	(0.934)	5.777	(1.227)	5.802	(1.097)	0.083
Observations	2790		2542		5332		

Table 7: Results for 3.5% discontinuity sample with bootstrapped samples of size 62. Standard deviations are in parentheses and the p-value corresponds to a t-test for the difference between the means in the group above and below the threshold.

[34] This could be due to the fact that less senior principals select themselves in schools with a lower percentage of disadvantaged students.

	Above Cutoff		Below Cutoff		Full Sample		p-value
Retention	0.030	(0.170)	0.042	(0.201)	0.036	(0.186)	0.025
Gender	0.497	(0.500)	0.461	(0.499)	0.479	(0.500)	0.015
Special Needs	0.000	(0.021)	0.001	(0.037)	0.001	(0.030)	0.309
Teacher Age	4.022	(0.333)	4.023	(0.260)	4.022	(0.299)	0.955
Teacher Seniority	3.867	(0.452)	3.932	(0.330)	3.899	(0.398)	0.000
Teacher Training	0.982	(0.025)	0.983	(0.026)	0.983	(0.026)	0.805
Principal Age	6.022	(1.308)	6.000	(1.206)	6.011	(1.259)	0.556
Principal Seniority	5.778	(1.228)	5.818	(0.912)	5.798	(1.083)	0.212
Observations	2250		2200		4450		

Table 8: Results for 3.7% discontinuity sample with bootstrapped samples of size 50. Standard deviations are in parentheses and the p-value corresponds to a t-test for the difference between the means in the group above and below the threshold.

	Above Cutoff		Below Cutoff		Full Sample		p-value
Retention	0.029	(0.168)	0.040	(0.196)	0.034	(0.182)	0.026
Gender	0.490	(0.500)	0.464	(0.499)	0.477	(0.500)	0.058
Special Needs	0.000	(0.019)	0.001	(0.038)	0.001	(0.030)	0.174
Teacher Age	4.022	(0.333)	4.023	(0.260)	4.022	(0.299)	0.950
Teacher Seniority	3.867	(0.452)	3.932	(0.330)	3.899	(0.398)	0.000
Teacher Training	0.982	(0.025)	0.983	(0.026)	0.983	(0.026)	0.784
Principal Age	6.022	(1.308)	6.000	(1.206)	6.011	(1.259)	0.512
Principal Seniority	5.778	(1.227)	5.818	(0.912)	5.798	(1.083)	0.165
Observations	2790		2728		5518		

Table 9: Results for 3.7% discontinuity sample with bootstrapped samples of size 62 (the smallest school in the sample). Standard deviations are in parentheses and the p-value corresponds to a t-test for the difference between the means in the group above and below the threshold.

With respect to the results of the BCF-IV algorithm, when we increase the number of sampled units we report the results just for the outcome variable *Progress School*, for which we find evidence of significant treatment effect variation. The main differences between the results for the sample of 50 students (Figure 4) and the ones for the sample of 62 students (Figure 8) are the following: (i) the overall effect is positive (but, again, not statistically significant); (ii) the conditional effects estimated are larger and with a higher statistical significance (which is not surprising given the larger sample).

Figures 9 and 10 depict the results obtained for the estimation of the Intention-To-Treat (ITT) effect using the same set-up used for the estimation of CACE in Figures 3 and 4. Figure 9 depicts the results for the ITT for the *A-certificate* outcome and can be directly compared with Figure 3, while Figure 10 depicts the results for the ITT for the *Progress School* outcome and can be directly compared with Figure 4. The results for CACE and ITT are fairly similar as the complier sub-populations vary between a maximum of 83% of compliers to a minimum of 58% of them. Hence, we can argue that our algorithm catches treatment variations in CACE that are directly driven by the effect variations in the estimated ITT and not by variations in the proportion of compliers within the different sub-populations.

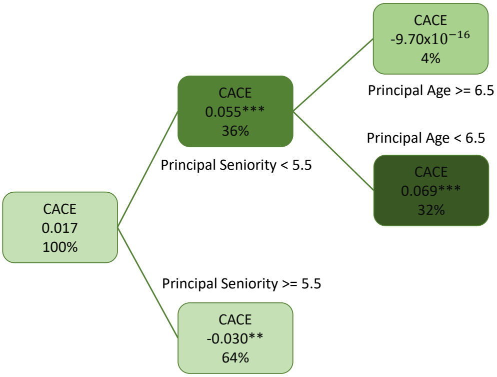


Figure 8: Visualization of the heterogeneous Complier Average Causal Effects (CACE) of additional funding on *Progress School* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in P. R. Hahn et al., 2020. The overall sample size is 5,518. The significance level is * for a significance level of 0.1, ** for a significance level of 0.05 and *** for a significance level of 0.01.

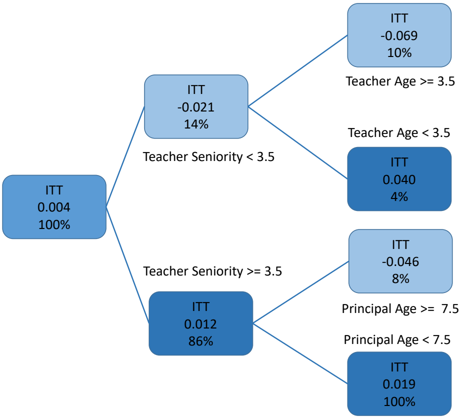


Figure 9: Visualization of the heterogeneous Intention-To-Treat (ITT) effect of additional funding on *A-certificate*. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in P. R. Hahn et al., 2020. The overall sample size is 4,300.

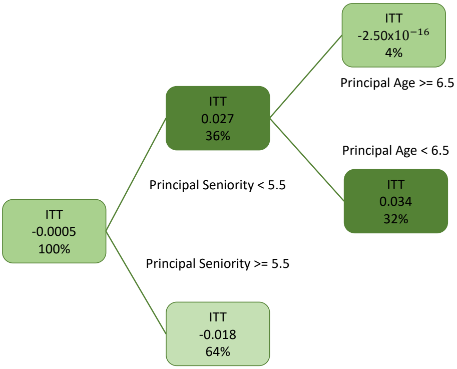


Figure 10: Visualization of the heterogeneous Intention-To-Treat (ITT) effect of additional funding on *Progress School* estimated using the proposed model. The tree is a summarizing classification tree fit to posterior point estimates of individual treatment effects as in P. R. Hahn et al., 2020. The overall sample size is 4,450.

Chapter 2

Causal rule ensemble: interpretable inference of heterogeneous treatment effects

There have been many developments in estimating the average treatment effects (ATE). In various fields, the estimation of the ATE provides a central insight on the causal effect of a treatment (e.g., a drug, a public intervention, an environmental policy, and so on) on an outcome, on average for the whole population. However, in addition to the ATE, it is critically important to identify subgroups of the population that would benefit the most from a treatment and/or would be most vulnerable to an environmental exposure. In the context of air pollution, it is deemed important to public health to identify the subgroups that are most vulnerable, so that effective interventions can be put in place to mitigate adverse health effects (K. Lee et al., 2018).

There is extensive literature on assessing heterogeneity of causal effects that is based on estimating the conditional average treatment effect (CATE). For each combination of covariates $\mathbf{X} = \mathbf{x}$ (i.e., subset of the

This Chapter is based on Lee, Bargagli-Stoffi and Dominici (2020).

features space), the average treatment effect conditioning on x can be estimated with the same set of the causal assumptions that are needed for estimating the ATE (Athey & Imbens, 2016). For example, earlier works on nonparametric approaches for estimating CATE relies on nearest-neighbor matching and kernel methods (Crump et al., 2008; M.-j. Lee, 2009). Wager and Athey (2018) discuss that these approaches may fail in handling a large number of covariates. This issue is often referred to as *curse of dimensionality* (Bellman, 1961; Robins & Ritov, 1997). Recently, other nonparametric approaches have been developed that rely on machine learning methods such as Random Forest (Breiman, 2001) and Bayesian Additive Regression Tree (BART) (Chipman et al., 2010). These approaches have been successful even when the number of features is very large. For instance, in their seminal contributions Foster et al. (2011) and Hill (2011) used forest-based algorithms for the prediction of the missing potential outcomes. In a similar spirit, P. R. Hahn et al. (2020) proposed a BART-based approach but with a novel parametrization of the outcome surfaces. In more recent contributions, Wager and Athey (2018) and Athey et al. (2019) developed forest-based methods for the estimation of heterogeneous treatment effects. They also provide asymptotic theory for the conditional treatment effect estimators and valid statistical inference.

Despite the success in accurately estimating the CATE using machine learning methods, these tree ensemble methods offer little guidance about which covariates or, even further, subgroups (i.e., subsets of the features space defined by multiple covariates) play an important role in driving treatment effects heterogeneity. Their parametrization of the covariate space is complicated and difficult to interpret by human experts. This issue is well-known as *lack of interpretability*. Increasing model interpretability is key to understanding and furthering human knowledge. However, effort to improve interpretability is so far lacking in the current causal inference literature dealing with the study of treatment effect heterogeneity.

As we show in Chapter 1 binary trees provide a high level of interpretability but suffer of three main limitations: (i) they are prone to overfitting data, so the results can fail to be replicable/generalizable (Strobl et al., 2009); (ii) they are able to discover just a limited number of heterogeneous subpopulations; (iii) they may fail to distinguish between effect modifiers and confounders.

In this Chapter, we propose a novel Causal Rule Ensemble (CRE) method that ensures interpretability of the causal rules, while maintaining a high level of accuracy of the estimated treatment effects for each rule, and accommodating for the shortcomings of tree-based methodologies. We define interpretability as the degree to which a human can understand the cause of a decision or consistently predict the results of the model (B. Kim et al., 2016; T. Miller, 2019). Decision rules are ideal for this non-mathematical definition of interpretability. A decision rule consists of a set of conditions about the covariates, that define a subset of the features space, and corresponds to a specific subgroup. The concept of interpretable causal rules was first introduced in the causal inference literature in the seminal contribution of T. Wang and Rudin (2017). However, these authors focus on a small set of short rules to capture the heterogeneous subgroups. In this Chapter, we improve over this paper by discovering a potentially larger and more complex set of rules, while maintaining a high level of interpretability.

We want to achieve at least three main goals: to (1) discover de novo the rules that lead to heterogeneity of causal effects; (2) make valid and precise inference about the CATE with respect to the newly discovered rules; and (3) assess sensitivity to unmeasured confounding bias for the rule-specific causal effects. To do so, we follow Athey and Imbens (2016), and we rely on a sample-splitting approach that divides the total sample into two smaller samples: (1) one for discovering a set of interpretable decision rules that could lead to treatment effect heterogeneity (i.e., discovery sample) and (2) the other for estimating the rule-specific treatment effects and associated statistical uncertainty (i.e., inference sample). We also tailor a sensitivity analysis method proposed by Q. Zhao et al. (2019) to assess the robustness of the rule-specific treatment effects to unmeasured confounding.

To summarize, our proposed approach has the following characteristics all embedded within a single algorithm: 1) identifies de novo the features that lead to treatment effect heterogeneity; 2) ensures that these features are interpretable, that is, they are mapped into simple causal rules that describe subsets of the study population where the treatment effect for that subgroup deviates from the ATE; 3) estimates the causal treatment effects for each of the rule, while maintaining statistical accuracy; 4) allows the researcher to detect, in a more flexible and stable way, a larger set of causal rules with respect to the one that could be discovered with state-of-the-art methodologies; 5) provides a powerful tool to disentangle the effect modifiers (namely, drivers of causal effect heterogeneity) from the confounders; 6) allows for a sensitivity analysis to unmeasured confounding bias; 7) provides higher standards of replicability.

The reminder of the Chapter is organized as follows. In Section 1, we introduce the main definitions of CATE and interpretable decision rules. In Section 2, we describe the algorithm used for the discovery of causal rules. Section 3 introduces our innovative ideas regarding estimation and sensitivity analysis. In Section 4, we conduct simulations studies. In Section 5, we apply the proposed method to the Medicare Data. Section 6 discusses the strengths and weaknesses of our proposed approach and areas of future research.

1 Treatment effect heterogeneity, interpretability and sample-splitting

1.1 Causal treatment effects

Suppose that we have N subjects. Following the notation and the potential outcome framework (Neyman, 1923, 1990; Rubin, 1974) introduced in Chapter 1, we can define, again, two potential outcomes $Y_i(1)$ and $Y_i(0)$ as a function of the treatment W_i assigned to each subject i (i.e., $Y_i(W_i)$). $Y_i(1)$ is the potential outcome for unit i under treatment, while $Y_i(0)$ is the potential outcome under control. As in Chapter 1, we consider the conditional average treatment effect (CATE) $\tau(x)$ defined as:

$$\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0)|\mathbf{X}_i = x\right], \tag{2.1}$$

and the average treatment effect (ATE) defined as $\tau = \mathbb{E}[\tau(x)]$.

Although $\tau(x)$ cannot be observed, $\tau(x)$ can be estimated under the assumption of *unconfoundedness* introduced in Chapter 1:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp W_i \mid \mathbf{X}_i. \tag{2.2}$$

This assumption means that the two potential outcomes depend on $\mathbf{X}_i$, but are independent of W_i conditioning on $\mathbf{X}_i$. By using the propensity score $e(x) = \mathbb{E}[W_i|\mathbf{X}_i = x]$ (Rosenbaum & Rubin, 1983b), the CATE $\tau(x)$ can be identified as

$$\tau(x) = \mathbb{E}\left[\left(\frac{W_i}{e(x)} - \frac{1 - W_i}{1 - e(x)}\right) Y_i \mid \mathbf{X}_i = x\right]. \tag{2.3}$$

However, in practice we do not know whether a considered set $\mathbf{X}_i$ is sufficient for the assumption (2.2). When there exists a source of unmeasured confounding, this assumption is violated, and the identification results do not hold. Sensitivity analysis provides a useful tool to investigate the impact of unmeasured confounding bias, which will be discussed in Section 3.2.

1.2 Interpretability and decision rules

Our primary goal is to provide an interpretable structure of $\tau(x)$ (discovery of treatment effects heterogeneity) and then estimate $\tau(x)$ in an efficient and precise manner (estimation of treatment effects heterogeneity). The functional form of $\tau(x)$ is not known, and may have a complicated structure varying across subgroups. To get a better estimate of $\tau(x)$, a complex model can be used, but such model may mask the informative structure of $\tau(x)$. Instead, it is possible to "extract" (or "approximate") important subspaces of $\mathbf{X}$ that are responsible for the variation of $\tau(x)$. The extracted subspaces may not perfectly describe the treatment effect heterogeneity, but provide a concise and informative summary of

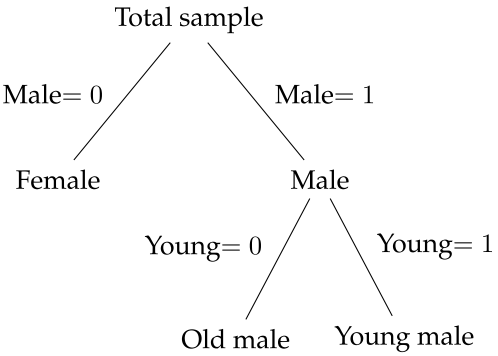


Figure 11: Example of causal rules.

Table 10: Extracting rules from the example tree in Figure 11

Rules	Conditions
r_1	Male= 0
r_2	Male= 1
r_3	Male= 1 & Young= 0
r_4	Male= 1 & Young= 1

$\tau(x)$. This is often referred to as *trade-off between interpretability and accuracy*. More specifically, weighing too much on achieving high accuracy in the estimation of the causal effect for a given covariates-profile generally comes at the cost of compromising interpretability. On the other hand, trying to achieve higher levels of interpretability might lead to the need to specify more complex models with a larger number of parameters, which in turn may reduce the statistical precision.

Our work focuses on how to find a good balance between interpretability and estimation accuracy. Even better, we aim to improve interpretability while the level of accuracy is maintained. To do so, we consider decision rules as base learners, and describe the heterogeneous treatment effect as a linear combination of these learners.

First, we define decision rules with formal definition. Let S_p be the set of possible values of the pth covariate and $s_{p,m} \subseteq S_p$ be a specific subset corresponding to the mth rule. Then, each decision rule $r_m(x)$ can be represented as

$$r_m(x) = \prod_{p:s_{p,m} \neq S_k} I(x_k \in s_{p,m}). \tag{2.4}$$

Define the covariate space $D = S_1 \times \cdots \times S_P$ as a Cartesian product of P sets. A vector of covariates $\mathbf{X}_i$ must lie in D for all i. Also, we define the subset D_m corresponding to the rule $r_m(x)$, $D_m = s_{1,m} \times \cdots \times s_{P,m}$. Then, $r_m(x) = 1$ if $x \in D_m$ and $r_m(x) = 0$ otherwise. Decision rules are easily obtained from a decision tree. For example, Figure 11 shows a toy example. The decision tree in this figure consists of several decision rules. Four decision rules can be extracted from this tree. For example, the young male group can be expressed as $r_4 = 1(\text{Male} = 1) \times 1(\text{Young} = 1)$. Other decision rules are listed in Table 10. We note that r_2 represents the internal node of the male group. Each decision rule corresponds to either internal or terminal nodes (leaves) except for the initial node (root node). Including rules corresponding to internal nodes may increase the total number of decision rules to consider, but it can be helpful not to miss important decision rules that are essential to describe variability of $\tau(x)$.

1.3 Sample Splitting

Analysis of heterogeneous treatment effects or subgroup analysis are typically conducted for subgroups defined *a priori* to avoid the cherry-picking problem that reports only subgroups with extremely high/low treatment effects (Cook et al., 2004). However, defining subgroups a priori requires a fairly good understanding of the treatment effect, probably from previous literature. On top of that, another problem is that researchers may miss unexpected subgroups. To overcome these limitations, we propose to use data-driven machine learning approaches combined with honest inference (Athey & Imbens, 2016). In particular, we use a sample-splitting approach that divides the total sample into two smaller samples (Athey & Imbens, 2016; K. Lee et al., 2018): (1) discovery and (2) inference samples. By using the discovery sample, decision rules are generated, and among them, a few candidates are selected.

Algorithm 1 Overview of the Causal Rule Ensemble (CRE) Method

```
• Randomly split the total sample into two smaller samples: the discovery and inference subsamples
• The Discovery Step (performed on the discovery subsample):
    1. Rules generation (Section 2.1)
        (a) Estimate $\tau_i$ using existing methods for estimating the CATE
        (b) Use $(\hat{\tau}_i, \mathbf{X}_i)$ to generate a collection of trees through tree-ensemble methods
        (c) From the collection, extract decision rules $r_j$by removing duplicates
    2. Rules regularization (Section 2.2)
        (a) Generate a new vector $\tilde{\mathbf{X}}_i^*$ whose $j$th component corresponds to rules $r_j$
        (b) Apply stability selection to $(\hat{\tau}_i, \tilde{\mathbf{X}}_i^*)$ and select potentially important decision rules $\tilde{\mathbf{X}}$.
• The Inference Step (performed on the inference subsample)::
    1. Estimate the CATE for the set of important decision rules selected in step 2.b (Section 3.1)
    2. Sensitivity analysis for the estimated CATE from the previous step (Section 3.2)
```

These selected rules are regarded as rules given a priori when making statistical inference using the remaining inference sample. This sample-splitting approach is transparent and furthermore even efficient in high-dimensional settings.

2 Discovery step: which subgroups are potentially important?

This Section illustrates the step for discovering potentially important subgroups in terms of decision rules. In particular, we introduce a generic method that creates a set of decision rules and identifies the rule-generated structure. The main advantage of separating this step from the inference step is to provide *transparency*. In this context, transparency refers to independence from one's subjective decisions. The procedure illustrated below is performed only using the data in the discovery sample.

2.1 Rule Generation

We use decision rules as base learners. Decision rules can be externally specified by using prior knowledge. However, we propose to employ a data-driven approach that uses hundreds of trees and extracts decision rules[1]. This approach overcomes two drawbacks that classical subgroup analysis approaches have: (1) they strongly rely on the subjective decisions on which are the heterogeneous subpopulations to be investigated; and (2) they fail to discover new rules other than the ones that are *a priori* defined by the researchers. On top of that, a data-driven method to detects causal rules avoids potential problems related to cherry-picking of the subgroups with extremely high/low treatment effects (Cook et al., 2004).

[1] Please note that these two approaches to decision rules discovery (i.e., prior knowledge vs data-driven discovery) can be merged in the proposed CRE algorithm. Indeed, one can manually tune the algorithm in order to let it discover more often subpopulations define by covariates that were found to be important in previous studies.

To create base learners for $\tau(x)$, we estimate τ_i first (where $\tau_i = tau(X_i)$). This estimation is not considered for making inference, but designed for exploring the heterogeneous treatment effect structure. Many different approaches for the estimation of $\tau(x)$ can be considered. There has been methodological advancement in directly estimating $\tau(x)$ by using machine learning methods such as Causal Forest (Wager & Athey, 2018) or Bayesian Causal Forest (BCF) (P. R. Hahn et al., 2020). The BCF method directly models the treatment effect such as $\tau_i = m_0(\mathbf{X}_i; \hat{e}) + \alpha(\mathbf{X}_i; \hat{e})W_i$ using the estimated propensity score $\hat{e}(\mathbf{X}_i)$. P. R. Hahn et al. (2020) shows that BCF produces extremely precise estimates of $\tau(x)$. This evidence is consistent with recent works showing an excellent performance of Bayesian machine learning methodologies in causal inference scenarios (Bargagli-Stoffi et al., 2019; P. R. Hahn, Dorie, et al., 2018; Hill, 2011; Logan et al., 2019; Nethery et al., 2019; Starling et al., 2019). The BCF method can be applied to the discovery sample to estimate τ_i. We denote such estimates with $\hat{\tau}_i^{BCF}$. Alternative approaches for the estimation of $\tau(x)$ are discussed in Appendix 8. Here, we want to highlight that the simulations' results show that the performance of the CRE algorithm in discovering the true causal rules is higher when we use BCF to estimate $\tau(x)$ (additional details on these simulations are provided in Appendix 8).

Once the unit level treatment effect $\hat{\tau}_i$ is obtained, one can fit a decision tree using the data $(\hat{\tau}_i, \mathbf{X}_i)$. After fitting a tree, decision rules can be extracted as discussed above. However, using a single tree is neither an efficient nor a stable way to find important decision rules. This is due to two main factors: (i) the *greedy* nature of tree-based algorithms and (ii) their lack of flexibility. Binary trees are greedy algorithms as they do not subdivide the population based on the overall best splits (the set of splits that would lead to the minimization of the overall criterion function), but they pick the best split at each step (the one that minimizes the criterion function at that particular step)[2]. Moreover, binary trees may not spot *simultaneous* patterns of heterogeneity in the data because of their binary

[2] Optimal trees (Bertsimas & Dunn, 2017) accommodate for this shortcoming at the cost of an extremely higher computational burden.

nature. Imagine that the treatment effect differs by sex and high school diploma. The tree-based algorithm may be able to spot only one of the two drivers of heterogeneity if one affects the treatment effect more than the other. Even when both the heterogeneity drivers are discovered they are spotted in a suboptimal way as an interaction between the two variables (i.e., women with a high school diploma, men with no diploma, etc).

To accommodate for these shortcomings of single trees, tree ensemble methods such as Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001) can be applied to obtain a collection of trees. Nalenz and Villani (2018) discuss that boosting and Random Forest are very different in nature, thus using both approaches can generate a wider set of decision rules. Even after removing duplicate rules, this will lead to a larger and more diverse set of candidate rules, but will increases the probability to capture important rules. We follow Friedman and Popescu (2008) and Nalenz and Villani (2018) to use the same settings of the tuning parameters for gradient boosting and Random Forest.

Our approach puts more attention on creating a large enough set that includes important decision rules with high probability. The created set will be reduced by using variable selection techniques later in the same discovery step. Another remark is that one should avoid using too lengthy (i.e., many conditions) or too many decision rules, which results in reducing interpretability.

2.2 Rules regularization and stability selection

Denote $r_m(x)$ as the generated rules from Section 2.1, with $m = 1, \dots, M^*$. Since each rule $r_m(x)$ indicates whether x satisfies the rule or not, it can take a value either 0 or 1. Define $\tilde{\mathbf{X}}^*$ as a new matrix whose columns are the decision rules. The number of rules, M^*, is usually larger than P and depends on how heterogeneous $\tau(x)$ is. Although the original dataset $\mathbf{X}$ is not high-dimensional, $\tilde{\mathbf{X}}^*$ can be high-dimensional.

The set of the generated rules is a set of potentially important rules. It can contain actually important decision rules that describe the true heterogeneity in the treatment effects. However, insignificant rules may be contained in the set. Then, rule selection is applied to distinguish between few important rules and many insignificant ones. This regularization step improves understanding of the heterogeneity in the effects and increases efficiency. On top of this, such step improves interpretability: if too many rules are selected, this information may be too complex to understand. We consider the following linear regression model of the form,

$$\tau(x) = \beta_0 + \sum_{m=1}^{M^*} \beta_m r_m(x) + \epsilon \tag{2.5}$$

Since a linear model is considered, the model (2.5) lends a familiar interpretation of the coefficients $\{\beta_m\}_0^{M^*}$. Using this linear model, one can employ the following penalized regression to select important rules:

$$\underset{\beta_m}{\operatorname{argmin}} \left\{ \beta_0 + \sum_{m=1}^{M^*} \beta_m r_m(x) \right\} \text{ subject to } ||\beta_m||_p \leq \lambda \tag{2.6}$$

where λ is the regularization parameter and $|| \cdot ||_p$ is the l_p-norm. However, variable selection has been known as a notoriously difficult problem.

The Least Absolute Shrinkage and Selection Operator (LASSO) estimator (Tibshirani, 1996) has been popular and widely used over the past two decades in order to solve the problem in (2.6) by employing the l_1-norm ($\sum_{m=1}^{M^*} |\beta_m|$). The usefulness of this estimator among other penalization regression methods is demonstrated in various applications (Belloni, Chernozhukov, Hansen, & Kozbur, 2016; Chernozhukov et al., 2017; Chernozhukov et al., 2016; L. Su et al., 2016). Using a regularization parameter λ, we can see that the estimate $\hat{\boldsymbol{\beta}}$ shrinks toward to zero as λ increases, which provides a sparse structure close to the true model. Consistency of LASSO variable selection has been studied in P. Zhao and Yu (2006) and others. However, the biggest challenge is to choose a proper value of λ for consistent selection. Cross-validation is usually accompanied to choose λ. However, it may fail for high-dimensional data (Meinshausen & Bühlmann, 2006). Stability selection (Meinshausen & Bühlmann, 2010) can be used to enhance the performance of the LASSO estimator. Roughly speaking, by using a subsampling scheme, selection probabilities of variables (decision rules) can be estimated. Given two parameters (i.e., cut-off threshold and the average number of selected variables), variable selection can be done in a comparatively robust and transparent way. Also, Meinshausen and Bühlmann (2010) discussed that the solution of stability selection depends little on the initial regularization chosen, which is a desirable feature when selecting decision rules.

Among initially generated M^* decision rules, we assume that M (with $M \ll M^*$) decision rules are selected as an output from the discovery procedure. Then, we can define $\tilde{\mathbf{X}}$ as the matrix containing only the selected decision rules. Figure 12 depicts the intuition behind these steps of rules discovery and selection. The Figure shows a simple forest composed of just five trees. Each node of each tree (with the exclusion of the roots) represents a causal rule (*rules' discovery*), while the nodes highlighted in red represent the causal rules that are selected by the stability selection methodology (*rules' selection*).

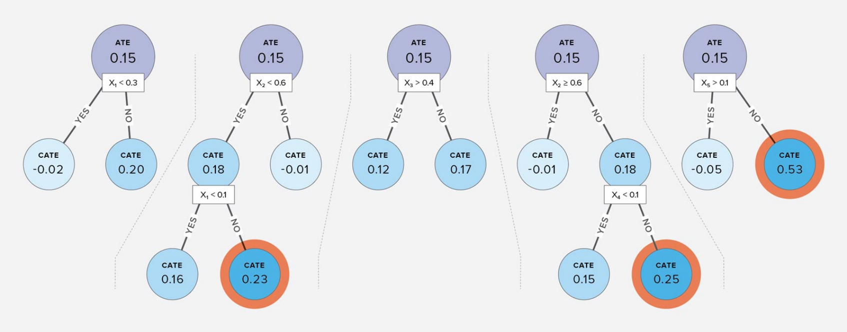


Figure 12: Rules' discovery and selection in a simple forest.

[2] The intuition behind stability selection is to aggregate the results from multiple penalized regressions constructed using sub-sampling techniques, controlling for a fixed number of false discoveries. Meinshausen and Bühlmann (2010) show that the variables that are selected more often are a *stable* set of important regressors.

3 Inference step: which subgroups are really different?

Using the rules that were discovered and then selected in the discovery step, the remaining inference sample is used to make inference. We propose a generic approach to estimate the rule-specific treatment effect, and compare it with existing approaches. Furthermore, we propose a new sensitivity analysis method to assess the presence of unmeasured confounding bias.

3.1 Estimating the subgroup-specific treatment effect

Decision rules are selected through the discovery step by using the linear model in (2.5),

$$\boldsymbol{\tau} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\mathbb{E}(\boldsymbol{\epsilon}|\mathbf{x}) = 0$ and $\text{var}(\boldsymbol{\epsilon}|\mathbf{x}) = \sigma^2 \boldsymbol{I}$. The main goal of this inference step is to estimate $\boldsymbol{\beta}$ that represents rule-specific treatment effects. We estimate the conditional average treatment effects for the subpopulations corresponding to the selected rules. We use ordinary least squares (OLS) to estimate $\boldsymbol{\beta}$. However, this estimator cannot be obtained as the unit level treatment effect $\boldsymbol{\tau}$ is not known. Hence, we define a new vector $\boldsymbol{\tau}^* = (\tau_1^*, \ldots, \tau_N^*)^T$ that is an estimate of $\boldsymbol{\tau}$. The newly defined $\boldsymbol{\tau}^*$ is just an intermediate value used in the inference step. Although $\boldsymbol{\tau}^*$ is an estimate, to distinguish this with the fitted value obtained by using the linear model, we save hat notation.

The estimate τ_i^* can be viewed as an observable quantity with some sampling error although τ_i^* is an estimated value. The quantity τ_i^* can be represented by

$$\tau_i^* = \tau_i + u_i \quad \text{where} \quad \mathbb{E}(u_i|\mathbf{x}_i) = 0 \text{ and } \text{var}(u_i|\mathbf{x}_i) = \omega_i. \tag{2.7}$$

In general, the variance ω_i is not constant across all individuals. By combining it with the model (2.5), the modified linear model is obtained as

$$\tau_i^* = \beta_0 + \sum_{j=1}^{M} \beta_j \tilde{X}_{ij} + \nu_i = \tilde{\mathbf{X}}_i \boldsymbol{\beta} + \nu_i, \tag{2.8}$$

where $\tilde{\mathbf{X}}_i$ is the ith row of $\tilde{\mathbf{X}}$ and $\nu_i = \epsilon_i + u_i$. Lewis and Linzer (2005) considered a similar linear model like the model (2.8) and found via simulation studies that the OLS estimator performs well in many cases. Also, they found that since the error ν_i is not homoscedastic, considering White or Efron's heteroscedastic-consistent standard error can provide an efficient estimate. Based on these findings, we also consider the OLS estimator for $\boldsymbol{\beta}$ that can be defined as

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\tau}^*. \tag{2.9}$$

Also, the fitted value $\hat{\boldsymbol{\tau}}$ is defined as $\hat{\boldsymbol{\tau}} = \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\tau}^*$.

The intermediate vector $\boldsymbol{\tau}^*$ can be chosen in various ways. However, to validate our estimator (2.9), each τ_i^* has to be unbiased with finite variance. Given that the matrix $\tilde{\mathbf{X}}$ is fixed, we can prove that the estimator $\hat{\beta}_j, j = 1, \ldots, M$ is a consistent estimator of β_j that is the average treatment effect for the subgroups defined by the decision rule r_j if r_j is not included in another rule $r_{j'}$.

Theorem 1. *If τ_i^* satisfies the model* (2.7) *(Condition 1) and $\mathbb{E}(\tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i) = \mathbf{Q}$ is a finite positive definite matrix (Condition 2), then the estimator $\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\tau}^*$ is a consistent estimator for $\boldsymbol{\beta}$.*

Corollary 1.1. *The estimator $\hat{\boldsymbol{\beta}}^{IPW} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\tau}^*$ where $\tau_i^* = \hat{\tau}_i^{IPW}$ is consistent*[3].

We need additional assumptions to prove asymptotic normality of $\hat{\boldsymbol{\beta}}$. For a general covariate matrix, the following three conditions are required:

3. $\mathbb{E}(\tilde{\mathbf{X}}_{ij}^4) < \infty$;

4. $\mathbb{E}(\nu_i^4) < \infty$;

5. $\mathbb{E}(\nu_i^2\tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i) = \boldsymbol{\Omega}$ is a positive definite matrix.

Since $\tilde{\mathbf{X}}_{ij}$ is either 0 or 1 in our setup, Condition (3) is satisfied by design. The following theorem represents the asymptotic distribution of $\hat{\boldsymbol{\beta}}$.

[3] The estimator $\hat{\boldsymbol{\beta}}^{IPW}$ is introduced in more detail in Appendix 8.

Theorem 2. *If Conditions (1)-(5) hold, then*

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}) \quad as \quad N \to \infty$$

where $\mathbf{V} = \mathbf{Q}^{-1}\boldsymbol{\Omega}\mathbf{Q}^{-1}$.

The variance $\mathbf{V}$ usually has to be estimated. The variance-covariance matrix estimator $\hat{\mathbf{V}}_n = \hat{\mathbf{Q}}^{-1}\hat{\boldsymbol{\Omega}}\hat{\mathbf{Q}}^{-1}$ can be obtained by the sandwich formula where $\hat{\mathbf{Q}} = n^{-1}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i$, $\hat{\boldsymbol{\Omega}} = n^{-1}\sum_{i=1}^{N}\hat{\nu}_i^2\tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i$ and $\hat{\nu}_i = \tau_i^* - \tilde{\mathbf{X}}_i\hat{\boldsymbol{\beta}}$. This estimator is robust and often called the White's estimator (H. White, 1980). There are other approaches to obtain a heteroscedasticity consistent covariance matrix, which is discussed in Long and Ervin (2000). For small samples, Efron's estimator (Efron, 1982), known as HC3 estimator, can be considered alternatively. Also, if the variance w_i is known from the large sample properties of existing methods for obtaining τ_i^*, then feasible generalized least squares estimators (Lewis & Linzer, 2005) can be considered.

Instead of estimation, hypothesis testing for identifying true decision rules can be considered. As we have seen in the previous simulation study, true rules representing treatment effect heterogeneity are discovered and selected with high probability. However, at the same time, non-true rules are selected with high probability. If one wants to know which rules are really important for describing treatment effect heterogeneity, variable selection can be used to choose a final model for treatment effect heterogeneity. For instance, the null hypothesis $H_0 : \boldsymbol{\beta} = 0$ can be tested by using a Wald-type test statistic $T_n = n\hat{\boldsymbol{\beta}}^T\hat{\mathbf{V}}_n^{-1}\hat{\boldsymbol{\beta}}$. If $T_n > \chi^2_{M,1-\alpha}$, H_0 is rejected.

3.2 Sensitivity analysis

The validity and consistency of the estimator $\hat{\boldsymbol{\beta}}$ rely on the assumption of no unmeasured confounders (2.2) and correct specification of the propensity score model. However, in reality, there is no guarantee that these assumptions are satisfied. Also, such assumptions are not directly testable. Without further knowledge about unmeasured confounding, it is extremely difficult to quantify how much bias can occur. Instead of attempting to quantify the degree of unmeasured confounding in the given dataset, it is more realistic to see how our causal conclusion will change with respect to various degrees of such bias. In this subsection, we propose a sensitivity analysis to examine the impact of potential violations of the no unmeasured confounding assumption on the discovered causal rules. We introduced a generic approach for estimating $\boldsymbol{\beta}$ in Section 3.1. However, in our sensitivity analysis, we consider the special case where τ_i^* is estimated as $\hat{\tau}_i^{SIPW}$[4]. Define $\mathbf{U} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T$ that is a $M \times N$ matrix. Also, let U_j be the jth row of $\mathbf{U}$ and U_{ji} be the (j,i) element of $\mathbf{U}$. Our estimator $\hat{\beta}_j$ is explicitly represented by

$$\begin{aligned}
\hat{\beta}_j &= \hat{\beta}_j(1) - \hat{\beta}_j(0) \quad \text{where} \\
\hat{\beta}_j(1) &= \left[\frac{1}{N}\sum_{i=1}^{N}\frac{W_i}{\hat{e}(\mathbf{X}_i)}\right]^{-1}\left[\sum_{i=1}^{N}\frac{U_{ji}Y_iW_i}{\hat{e}(\mathbf{X}_i)}\right] \\
\hat{\beta}_j(0) &= \left[\frac{1}{N}\sum_{i=1}^{N}\frac{1-W_i}{1-\hat{e}(\mathbf{X}_i)}\right]^{-1}\left[\sum_{i=1}^{N}\frac{U_{ji}Y_i(1-W_i)}{1-\hat{e}(\mathbf{X}_i)}\right].
\end{aligned}$$

We consider the marginal sensitivity model that was introduced by Tan (2006) and Q. Zhao et al. (2019). Let the true propensity probability $e_0(\mathbf{x}, y; a) = \mathrm{P}_0(W = 1|\mathbf{X} = \mathbf{x}, Y(a) = y)$ for $a \in \{0, 1\}$. If the assumption of no unmeasured confounders holds, this probability would be the same as $e_0(\mathbf{x}) = \mathrm{P}_0(W = 1|\mathbf{X} = \mathbf{x})$ that is identifiable from the data. Unfortunately, this assumption cannot be tested since $e_0(\mathbf{x}, y; a)$ is generally not identifiable from the data. For each sensitivity parameter Λ that will be introduced in detail later, the maximum deviation of $e_0(\mathbf{x}, y; a)$ from the identifiable quantity $e_0(\mathbf{x})$ is restricted, and sensitivity analysis is conducted for each Λ to see if there is any qualitative change of our conclusion. In addition to this non-identifiability of $e_0(\mathbf{x}, y; a)$, there is another difficulty in obtaining $e_0(\mathbf{x})$ non-parametrically when $\mathbf{X}$ is high-dimensional. In practice, as it is common use, $e_0(\mathbf{x})$ is estimated by a parametric logistic model in the form of $e_\gamma(\mathbf{x}) = \exp(\gamma'\mathbf{x})/\{1+\exp(\gamma'\mathbf{x})\}$

[4] The estimator $\hat{\boldsymbol{\beta}}^{SIPW}$ is introduced in more detail in Appendix 8.

where $e_{\gamma_0}(\mathbf{x})$ can be considered as the best parametric approximation of $e_0(\mathbf{x})$, and used for sensitivity analysis.

Our sensitivity model assumes that the true propensity probability $e_0(\mathbf{x}, y; a) = \mathrm{P}_0(W = 1|\mathbf{X} = \mathbf{x}, Y(a) = y)$ satisfies:

$$e_0(\mathbf{x}, y; a) \in \mathcal{E}_{\gamma_0}(\Lambda) = \{0 < e(\mathbf{x}, y; a) < 1 : 1/\Lambda \leq OR\{e(\mathbf{x}, y; a), e_{\gamma_0}(\mathbf{x})\} \leq \Lambda\} \tag{2.10}$$

for $a \in \{0, 1\}$ and where $e_{\gamma_0}(\mathbf{x}) = \mathrm{P}_{\gamma_0}(W = 1|\mathbf{X} = \mathbf{x})$ and $OR\{e(\mathbf{x}, y; a), e_{\gamma_0}(\mathbf{x})\}$ is an odds ratio:

$$OR\{e(\mathbf{x}, y; a), e_{\gamma_0}(\mathbf{x})\} = \frac{(1 - e(\mathbf{x}, y; a)) \cdot e_{\gamma_0}(\mathbf{x})}{e(\mathbf{x}, y; a) \cdot (1 - e_{\gamma_0}(\mathbf{x}))}. \tag{2.11}$$

The deviation of $e_0(\mathbf{x}, y; a)$ is symmetric with respect to the parametrically identifiable quantity $e_{\gamma_0}(\mathbf{x})$, and the degree of the deviation is governed by the sensitivity parameter $\Lambda \geq 1$. When $\Lambda = 1$, $e_0(\mathbf{x}, y; a) = e_{\gamma_0}(\mathbf{x})$ for all a, which implies that there is no violations of the assumptions. If the propensity score model is correctly specified, $e_{\gamma_0}(\mathbf{x}) = e_0(\mathbf{x})$. Also, if there is no unmeasured confounder, $e_0(\mathbf{x}, y; a) = e_0(\mathbf{x})$. Therefore, under the assumptions of correct propensity score model and no unmeasured confounder, $e_0(\mathbf{x}, y; a) = e_{\gamma_0}(\mathbf{x})$. Our sensitivity model considers violations of both assumptions. This sensitivity analysis model resembles the model proposed by Rosenbaum (2002). The connection between the two models is illustrated in Section 7.1 in Q. Zhao et al. (2019).

The $(1 - \alpha)$-coverage confidence interval of β_j can be constructed by using the percentile bootstrap. First, we replace $U_{ji}Y_i$ by $\tilde{Y}_i^j$ and treat $\tilde{Y}_i^j$ as if it is an observed outcome, then the confidence interval for each β_j can be constructed through the procedure in Algorithm 2. Confidence intervals have at least $100(1 - \alpha)$ % coverage probability even in the presence of unmeasured confounding. The validity of the percentile bootstrap confidence interval $[L_j, U_j]$ can be proved by using Theorem 1 in Q. Zhao et al. (2019). In Step (2e) of Algorithm 2, the optimization problem can be efficiently solved by separating two simpler optimization problems. The minimum is obtained when the first part $\frac{\sum_{i=1}^{N_1} \tilde{Y}_{ji}^{(\ell)}[1+q_i \exp\{-\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}{\sum_{i=1}^{N_1}[1+q_i \exp\{-\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}$ is minimized and the second part $\frac{\sum_{i=N_1+1}^{N} \tilde{Y}_{ji}^{(\ell)}[1+q_i \exp\{\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}{\sum_{i=N_1+1}^{N}[1+q_i \exp\{\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}$ is maximized. For instance, the minimization of the first part is achieved at $\{q_i : q_i = 1/\Lambda$ for $i = 1, \ldots, b,$ and $q_i = \Lambda$ for $i = b+1, \ldots, N_1\}$ for some b. To find the minimum, it is required to check every possible value of b, and the computational complexity is $O(N_1)$. See Proposition 2 in Q. Zhao et al. (2019) for more details.

Algorithm 2 Constructing the confidence interval of β_j for each sensitivity parameter Λ

```
1. Generate a matrix $\mathbf{U} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T$
2. In the $\ell$th of $L$ iterations:
    (a) Generate a bootstrapped sample: $(U_i^{(\ell)}, Y_i^{(\ell)}, \mathbf{X}_i^{(\ell)})_{i=1,\ldots,N}$.
    (b) Generate the transformed outcomes $\tilde{Y}_{ji}^{(\ell)} = U_{ji}Y_i$ for all $i$, where $U_{ji}$ is the $(j,i)$ element of $\mathbf{U}$.
    (c) Reorder the index such that the first $N_1 = \sum_{i=1}^{N} W_i^{(\ell)}$ units are treated with $\tilde{Y}_{j1} \geq \ldots \geq \tilde{Y}_{j,N_1}$ and the rest are control with $\tilde{Y}_{j,N_1+1} \geq \ldots \geq \tilde{Y}_{j,N}$
    (d) Compute the estimate $\hat{\gamma}^{(\ell)}$ by fitting the logistic regression with $(W_i^{(\ell)}, \mathbf{X}_i^{(\ell)})$
    (e) Solve the following optimization problems:
        $$\text{min or max} \frac{\sum_{i=1}^{N_1} \tilde{Y}_{ji}^{(\ell)}[1 + q_i \exp\{-\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}{\sum_{i=1}^{N_1}[1 + q_i \exp\{-\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]} - \frac{\sum_{i=N_1+1}^{N} \tilde{Y}_{ji}^{(\ell)}[1 + q_i \exp\{\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}{\sum_{i=N_1+1}^{N}[1 + q_i \exp\{\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}$$
        subject to $1/\Lambda \leq q_i \leq \Lambda$, for $1 \leq i \leq N$, and denote the minimum as $L_j^{(\ell)}$ and the maximum as $T_j^{(\ell)}$
3. Construct the $(1-\alpha)$-coverage confidence interval $[L_j, T_j]$ where $L_j = Q_{\alpha/2}\left(L_j^{(\ell)}\right)$ and $T_j = Q_{1-\alpha/2}\left(T_j^{(\ell)}\right)$.
```

4 Simulation

In this section, we introduce two simulation studies to assess the performance of the CRE method. In the first simulation study, we evaluate the discovery step in terms of how well the method performs in discovering the true underlying decision rules. In the second simulation study, we evaluate the overall performance of both the discovery and inference steps in terms of estimation accuracy.

4.1 Simulation study: rules discovery

In order to evaluate the ability of CRE in the discovery step we run a series of simulations in which we assess how many times CRE can spot the true underlying decision rules. First, we assess the *absolute* performance of CRE; then we compare its performance with the one of the Honest Causal Tree (HCT) method (Athey & Imbens, 2016). The HCT method is a causal decision tree algorithm that provides interpretable decision rules by discovering a set of causal rules that can be represented through a single binary tree. In order to evaluate the ability of discovering decision rules we consider, among the discovered rules, how many times true rules are captured.

As we discussed in Section 2, to apply the CRE method, many approaches can be considered to estimate the individual treatment effect τ_i. We have found via simulation studies that methods for directly estimating τ_i, such as the BCF method, have better performance than other approaches such as $\hat{\tau}_i^{IPW}$ or BART (in Appendix 8, we show the comparative performances of BCF, BART, IPW and outcome regression). Thus, in this simulation study, we implement the BCF approach for the CRE method. We call this version of the CRE method *CRE-BCF*. Also, for the data-generating process, we generate the covariate matrix $\mathbf{X}$ with 10 binary covariates from X_{i1} to $X_{i,10}$. For potential outcomes, we consider $Y_i(0) \sim N(X_{i1} + 0.5X_{i2} + X_{i3}, 1)$ and $Y_i(1) = Y_i(0) + \tau(\mathbf{X}_i)$ where $\tau = k$ if $X_{i1} = 0, X_{i2} = 0$, and $\tau = -k$ if $X_{i1} = 1, X_{i2} = 1$. The binary treatment indicator W_i is sampled from a binomial distribution, $W_i \sim \mathit{Binom}(\pi_i)$ where $\pi_i = \text{logit}(-1 + X_{i1} - X_{i2} + X_{i3})$. Finally, the output is generated as $y = y_0 \cdot (1 - w_i) + y_1 \cdot w_i + f(\mathbf{X})$ where $f(\mathbf{X})$ is a linear function of the confounders X_{i1}, X_{i2} and X_{i3}. Moreover, we introduce variations in (i) the number of decision rules, i.e. from two to four decision rules: $\tau = k$ if $X_{i1} = 0, X_{i2} = 0$ or $X_{i2} = 0, X_{i3} = 0$; $\tau = -k$ if $X_{i1} = 1, X_{i2} = 1$ or $X_{i2} = 1, X_{i3} = 1$; (ii) the effect size, i.e. $(k \cdot 0.1)_{k=0}^{20}$ with $k \in \mathbb{N}$; and (iii) the sample size: i.e. 1,000 and 2,000 data points.

Figure 13, depicts the simulations' results for the scenarios with two true rules (left panel) and four true rules (right panel). We consider 100 simulated datasets for each effect size and we report the average number of correctly discovered rules (CDR). We report the results for both 1,000 (red solid line) and 2,000 (red dashed line) data points. We find that in the scenario with 2,000 data points CRE-BCF is faster in discovering the actual rules. However, the difference in not sizable as the performance of CRE-BCF is excellent also with the smaller sample size. As the effect size k increases CRE-BCF is always able to spot the true causal rules. Even for smaller effect sizes, where the causal rules are not statistically significant (i.e., when the effect size is smaller than one), the performance of CRE is excellent both in the case with two and four true rules[5]. In Appendix 10, we run a series of simulations introducing additional variations in the correlation between the covariates and the functional form of $f(\mathbf{X})$. Here, it is worth highlighting that none of these variations in the data generating process decreases the ability of CRE to correctly spot the true underlying causal rules.

As we show that CRE-BCF has an excellent *absolute* performance in the discovery phase, it is interesting to compare the fit of this method with state-of-the-art methodologies for interpretable discovery of heterogeneous causal effects such as the HCT algorithm. Before comparing these two methods, it is important to highlight that the HCT algorithm is able to spot just those causal rules that can be represented through a single binary decision tree. The CRE-BCF method improves on this through its ability to spot a larger set of decision rules. Namely, the rules that can be detected through the CRE-BCF method are not just the rules that can be represented by a single binary tree, but any possible set of decision rules. In order to produce a more meaningful comparison between CRE-BCF and HCT, we restricted the data generating process to causal rules that are representable through a binary tree. In particular, in the case of

[5] In Appendix 9, we depict more detailed results for this set of simulations.

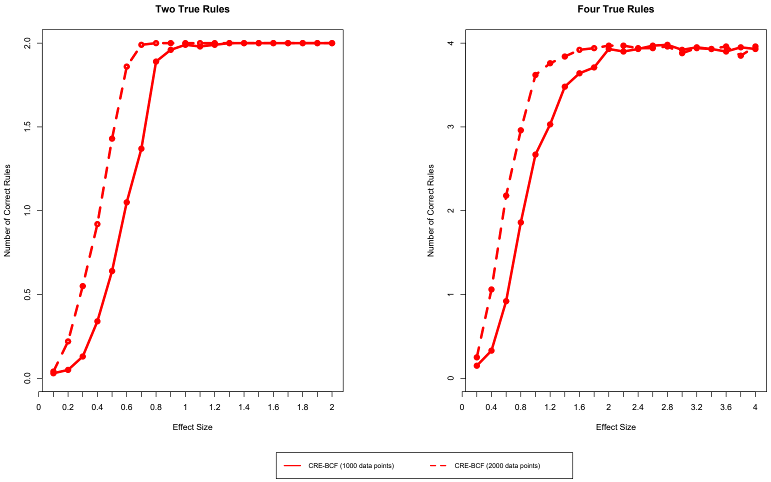


Figure 13: Average number of correctly discovered rules for CRE-BCF in the case of two true rules (left panel) and four true rules (right panel).

two causal rules we have that $\tau = k$ if $X_{i1} = 0, X_{i2} = 0$ and $\tau = -k$ if $X_{i1} = 0, X_{i2} = 1$; while in the case of four causal rules we have that $\tau = k$ if $X_{i1} = 0, X_{i2} = 0$, $\tau = 2k$ if $X_{i1} = 0, X_{i2} = 0, X_{i3} = 0$; $\tau = -k$ if $X_{i1} = 0, X_{i2} = 1$ and $\tau = -2k$ if $X_{i1} = 0, X_{i2} = 1, X_{i3} = 1$. This scenario was chosen because it is the most favourable to the HCT algorithm. Moreover, we introduce variations in the set of covariates that are used to define the causal rules (we refer to these variables as effect modifiers) by switching X_1 with X_{10}, X_2 with X_9 and X_3 with X_8. These scenarios where chosen to investigate whether or not there are differences between the two methodologies in settings where (i) the effect modifiers are the same covariates as the confounders; and (ii) the effect modifiers are different from the confounders. This is of central importance as confounders might affect the ability of the algorithm to spot the correct causal rules. Figures 14 and 15 depict the results in the case with the same variables for both confounders and effect modifiers and in the case of different variables, respectively.

In the former scenario, CRE-BCF consistently outperforms HCT,

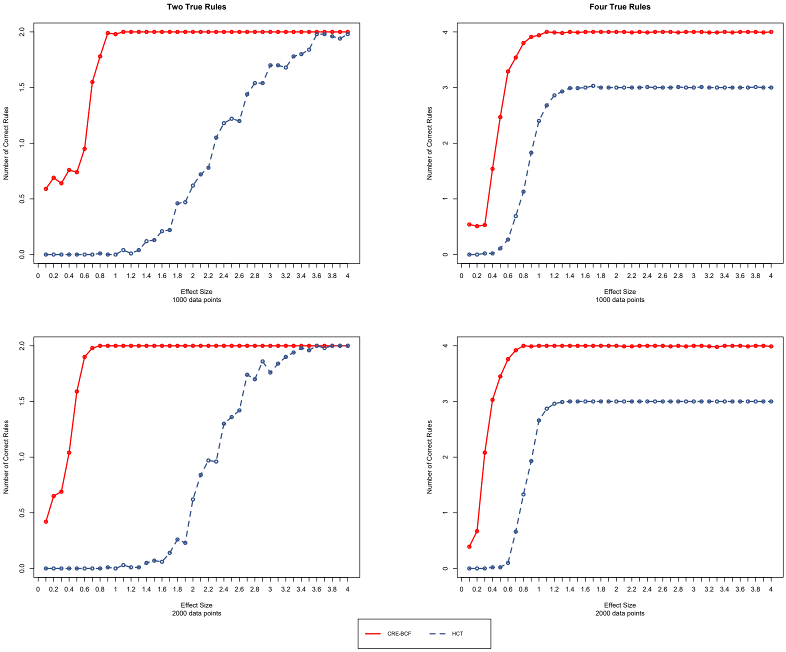


Figure 14: Average number of correctly discovered rules in the case where the same variables are used as confounders and effect modifiers. The first column depicts the case of two true rules while the second column the case of four true rules. In the first row the sample size is 1,000 while in the second it is 2,000.

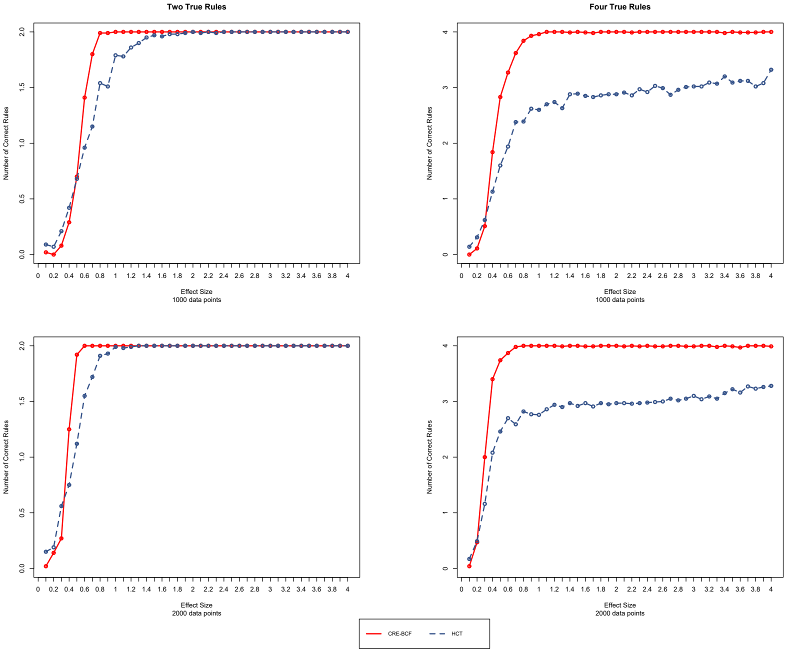


Figure 15: Average number of correctly discovered rules in the case where the confounders are different from the effect modifiers. The first column depicts the case of two true rules while the second column the case of four true rules. In the first row the sample size is 1,000 while in the second it is 2,000.

while in the latter scenario CRE-BCF is still performing better but the performance gap is narrower especially in the case with two true causal rules. In both the scenarios, the gap is wider as the number of true rules is four. This is due to the fact that HCT is consistently unable to detect all the four true rules. These simulations show that the usage of CRE-BCF results in an increase in the detection of the true causal rules especially when there is overlap between the confounders and effect modifiers. This is a very interesting feature of CRE-BCF as similar scenarios are very likely in real-world applications. For instance, it can be the case in pollution studies, that income is a confounder as poorer people live in neighbours with higher levels of pollution, and also an effect modifier as poorer people may have worst living conditions and, in turn, experience higher negative effects from pollution. On top of this, it is important to highlight that CRE provides a smaller number of detected rules (4 to 7) as compared to HCT (10 to 84).

4.2 Simulation study: rules effects estimation

In this subsection, we evaluate the overall performance of the CRE method including both the discovery and inference steps. In the previous simulation study, we found that the BCF approach shows a great performance in discovering underlying decision rules. Also, it has been shown that the BCF approach estimates τ_i with great accuracy heuristically. Thus, in this simulation study, we use the BCF approach for both discovery and inference steps when applying the CRE method. In particular, in the discovery step, $(\mathbf{X}_i, \hat{\tau}_i^{BCF})$ is used in order to discover and select important decision rules. In the inference step, the intermediate variable τ_i^* is estimated, as in the previous section, by using the BCF approach (CRE-BCF).

To compare, we consider an original BCF approach that does not use the sample-splitting technique, and we call this *original-BCF*. The original BCF provides the estimate $\hat{\tau}_i$, but does not provide an interpretable form of the heterogeneous treatment effects. On the contrary, the CRE-BCF not only provides a set of decision rules that significantly increase

Table 11: RMSE comparison between the CRE and BCF methods

	Method						
N	CRE50	CRE40	CRE30	CRE25	CRE20	CRE10	BCF
500	0.568	0.520	0.486	**0.478**	0.498	0.598	0.391
1000	0.399	0.366	0.347	**0.343**	0.344	0.406	0.355
1500	0.326	0.299	0.279	**0.276**	0.278	0.320	0.309
2000	0.283	0.258	0.239	**0.231**	0.234	0.262	0.237

interpretability of the finding, but also provides the estimate with respect to the discovered structure. It is difficult to compare the two methods in terms of interpretability, so in this simulation study, we compare them in terms of estimation accuracy. Athey and Imbens (2016) recommends to use a (50%, 50%) ratio between the discovery and inference samples for sample-splitting, but K. Lee et al. (2018) shows through simulation studies that a (25%, 75%) ratio has better performance that (50%, 50%). We investigate six different ratios from (10%, 90%) to (50%, 50%). Note that the original-BCF can be considered as the (0%, 100%) ratio of the CRE-BCF.

For the data generating process, covariates, treatment, and potential outcomes are generated in the same way as in the previous simulation study. In this simulation, only difference is that we assume that there are two true underlying decision rules: (1) $X_{i1} = 0, X_{i2} = 0$ and (2) $X_{i1} = 1, X_{i2} = 1$. The treatment effect τ_i is defined as $\tau_i = 1$ if $X_{i1} = 0, X_{i2} = 0$, $\tau_i = -1$ if $X_{i1} = 1, X_{i2} = 1$, and $\tau_i = 0$ otherwise. We consider four samples sizes, $N = 500, 1000, 1500$ and 2000. We consider the root mean squared error (RMSE) to compare the two methods.

Table 11 shows the performance comparison using RMSE. We consider 1000 simulated datasets for each sample size and provide the average of 1000 RMSE values. Among the considered ratios, (25%, 75%) provides the least RMSE for every sample size. The RMSE value decreases as the proportion of the discovery sample increases up to 25%, and it starts to increase after 25%. When the sample size is small (i.e., $N = 500$), the RMSE for CRE25 is higher than that for BCF, however, it is lower when N is moderately large. This simulation result shows that even though, for instance, the CRE25 method uses 75% of the total sample for inference, it is as efficient as the original BCF that uses 100% of the total sample for inference.

5 Data application: Medicaid data

We apply the proposed CRE method to the Medicare data in order to study the effect of long-term exposure to fine particulate matter (PM$_{2.5}$) on 5-year mortality. K. Lee et al. (2018) studied the treatment effect of exposure to PM$_{2.5}$ with 110,091 matched pairs and discovered treatment effect heterogeneity in a single tree structure by using the HCT approach. However, as pointed out in their discussion, a single tree unavoidably contains subgroups that are not informative for describing treatment effect heterogeneity due to the nature of tree algorithms. In particular, they found six disjoint subgroups, but only three of them are informative to describe the treatment effect heterogeneity.. Also, as we discussed and showed in our simulation study, the HCT approach can be highly affected by sample-to-sample variations. However, in this specific study, K. Lee et al. (2018) showed that the discovered tree is stable by evaluating 1000 bootstrapped discovery samples.

For a brief overview of the matched Medicare data, the data contain Medicare beneficiaries in New England regions in the United States between 2000 and 2006. The treatment is whether the two-year (2000-2001) average of exposure to PM$_{2.5}$ is greater than 12 $\mu g/m^3$. The outcome is five-year mortality measured between 2002-2006. There are four individual level covariates - sex (male, female), age (65-70, 71-75, 76-80, 81-85, 86+), race (white, non-white), and Medicaid eligibility (eligible, non-eligible). Medicaid eligibility is considered as a variable indicating socioeconomic status. If an individual is eligible for Medicaid, it is highly likely that he/she has lower household income, thus we use this variable as a proxy for low income. In the matched data, these four variables are exactly matched. There are also 8 ZIP code-level or 2 county-level covariates, and they are fairly balanced between the treated and control groups, see Table 2 in K. Lee et al. (2018) for the covariate balance. Also, see Di

Table 12: Discovering decision rules and estimating the coefficients for the decision rules

Rules		Covariates			
		Individual		Indiv. + ZIP code	
#	Description	Est.	95% CI	Est.	95% CI
	Intercept	0.070	(0.054, 0.087)	0.077	(0.061, 0.094)
r_1	$1(\text{white} = 0)$	-0.008	(-0.027, 0.011)		
r_2	$1(65 \leq \text{age} \leq 75)$	-0.012	(-0.024, 0.000)	-0.009	(-0.021, 0.002)
r_3	$1(65 \leq \text{age} \leq 80)$	-0.027	(-0.045, -0.010)	-0.027	(-0.045, -0.010)
r_4	$1(65 \leq \text{age} \leq 85) \cdot 1(\text{Medicaid} = 0)$	-0.033	(-0.050, -0.016)	-0.031	(-0.048, -0.015)
r_5	$1(\text{hispanic \%} = 0) \cdot 1(\text{education} = 1)$			-0.019	(-0.038, 0.000)
r_6	$1(\text{hispanic \%} = 0) \cdot 1(\text{education} = 1)$ $\cdot 1(\text{population density} = 0)$			-0.045	(-0.067, -0.022)

et al. (2016) for general description about the Medicare data.

We apply the CRE method to the same discovery and inference samples that are split by using a (25%, 75%) ratio and contain 27,500 and 82,591 matched pairs respectively. Since two individuals in a matched pair share the same covariates, but experience different treatment values, the observed outcomes can be considered as two potential outcomes for a hypothetical individual that represents the corresponding matched pair. The difference between the two outcomes can be considered as an estimate of τ_i for matched pair i. However, since our outcome is binary, the estimate $\hat{\tau}_i^{Matching}$ can take only one of three possible values $\{-1, 0, 1\}$. This discrete feature is undesirable under the linear regression model (2.5). Instead, we ignore the matched structure and use $27500 \times 2 = 55000$ individuals in the discovery sample as if they were obtained independently. We use the logistic BART approach to estimate potential outcome functions $m_w(x) = \mathbb{E}[Y_i(w)|\mathbf{X}_i = x]$, $w = 0, 1$. Then, the estimate $\hat{\tau}_i^{BART} = \hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)$ is obtained.

First, we focus on the discovery step with the estimate $\hat{\tau}_i^{BART}$. We apply the CRE method with four individual-level covariates. Four decision rules, r_1, r_2, r_3, r_4, are discovered. Table 12 shows the descriptions for these rules on the left column. Based on this finding, the model $\tau(x) = \beta_0 + \sum_{j=1}^{4} \beta_j r_j(x)$ is considered for the later inference step. The first rule is for non-white people that is only 7% of the total population. The next three rules are defined by age, and r_2 is included in r_3. The intercept β_0 represents the treatment effect for the subgroup of Medicaid eligible white people aged between 81-85 and white people aged above 85. This subgroup corresponds to the single tree finding in K. Lee et al. (2018).

Furthermore, we extend the CRE method to the four individual-level and eight ZIP code-level covariates. The ZIP code-level covariates could not be considered in K. Lee et al. (2018) because pairs were matched exactly only on individual-level covariates. When applying the CRE method, five rules are discovered including the three previous rules, r_2, r_3, r_4 and two additional rules, r_4, r_5. The additional rules are defined by Hispanic population (10% above or not), Education (did not complete high school 30% above or not), and Population density (above the average or not). The rule r_4 means a subgroup of people living in areas where the proportion of hispanic is below 10% and the proportion of people who did not complete high school is above 30%. The rule r_5 is included in r_4, and means a subgroup of r_4 with the additional condition that the population density is below the average.

Before making inference with the discovered rules, we want to emphasize two aspects in the discovery step of the CRE method. First, since the CRE method discovers decision rules instead of a whole tree, only important subgroups (decision rules) are selected, and other subgroups are represented by the intercept. For example, K. Lee et al. (2018) discovered 6 subgroups, but only a few of them are informative for describing the treatment effect heterogeneity. Second, the discovered rules are stable. Being stable means that if another discovery sample is chosen, the discovered rules are hardly changed while a discovered tree varies with its size and terminal nodes. This robustness to sample-to-sample variation makes findings replicable.This feature is extremely significant because higher standards of reproducibility are deemed important in the context of open science.

Next, with respect to the discovered decision rules, we use the remaining 82,591 pairs in the inference sample to estimate the rule-specific treatment effects. For the first rule set $r_1, \ldots, r_4$, the corresponding coefficients are estimated and reported in Table 12. To obtain the estimates and

95% confidence intervals, we consider $\tau_i^* = \hat{\tau}_i^{SIPW}$ to claim the asymptotic normality and obtain the 95% confidence intervals using the asymptotic distributions. Causal Forest method (Athey et al., 2019) can be also considered since it guarantees the asymptotic normal distribution for $\boldsymbol{\beta}$. On the first part of the right column of Table 12, all the coefficients except for the intercept have the negative sign. The intercept indicates that the subgroup which does not belong to the discovered rules from r_1 to r_4 is significantly affected by exposure to air pollution. There is a 7 percentage point increase of mortality rate for this subgroup. Though the estimates for r_1, r_2 are negative, they are not statistically significant at a significance level $\alpha = 0.05$. However, the estimates for r_3, r_4 are significant, which means that people below 80 and people below 85 not eligible for Medicaid are significantly less vulnerable to air pollution exposure. When including ZIP code-level covariates, we found rules from r_2 to r_6. Similarly, all the estimates are negative and the rule r_2 is not significant. For newly discovered r_5, r_6, only r_6 is statistically significant.

Also, we want to emphasize the interpretation of the coefficients in the inference step. A non-significant estimate does not mean that the corresponding subgroup has the null treatment effect. In this context, non-significance means that the decision rule is not no longer important for describing the treatment effect heterogeneity. For instance, consider a subgroup of black people above 85 who are represented by only the rule r_1 that is shown as not significant. However, the treatment effect for this subgroup is $\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 = 0.062$ with the 95% CI (0.041, 0.084). Also, a subgroup of black people below 75 has the effect $\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\beta}}_3 = 0.023$ (95% CI: (0.003, 0.043)), which is still significant. Since we have the asymptotic distribution for $\boldsymbol{\beta}$, any subgroup's treatment effect can be examined.

Finally, to evaluate the robustness of the above finding about the treatment effect heterogeneity, we conduct sensitivity analysis in the inference step by setting $\tau_i^* = \hat{\tau}_i^{SIPW}$. We use the following model for sensitivity analysis, $\tau_i = \beta_0 + \sum_{j=2}^{6} \beta_j r_j$. Under the sensitivity model (2.10), for each Λ, we obtain several sets of 95% CIs for the coefficients. Table 13 shows the 95% CIs for Λ from 1.01 to 1.05. As Λ increases all the CIs get

Table 13: Sensitivity analysis for the treatment effect heterogeneity by using the percentile bootstrap

Rules	Sensitivity Parameter Λ				
	1.01	1.02	1.03	1.04	1.05
Inter.	(0.054, 0.100)	(0.045, 0.110)	(0.035, 0.121)	(0.024, 0.130)	(0.014, 0.140)
r_2	(-0.026, 0.007)	(-0.031, 0.012)	(-0.036, 0.017)	(-0.041, 0.022)	(-0.047, 0.028)
r_3	(-0.051, -0.003)	(-0.060, 0.004)	(-0.069, 0.014)	(-0.077, 0.023)	(-0.086, 0.033)
r_4	(-0.054, -0.009)	(-0.062, 0.000)	(-0.071, 0.009)	(-0.078, 0.017)	(-0.087, 0.024)
r_5	(-0.040, 0.004)	(-0.046, 0.012)	(-0.053, 0.016)	(-0.059, 0.020)	(-0.066, 0.026)
r_6	(-0.072, -0.017)	(-0.080, -0.011)	(-0.085, -0.004)	(-0.091, 0.002)	(-0.098, 0.008)

wider. When $\Lambda = 1.04$, all the coefficients contain zero and there is no evidence for the heterogeneity. In particular, $\Lambda = 1.04$ means that if there is an unmeasured confounder that can make the estimated propensity score deviate from the true score by 1.04 in terms of the odds ratio scale, then our finding about the heterogeneity can be explained by this unmeasured bias. One further thing to note is that even if the heterogeneity can be explained by an unmeasured bias at $\Lambda = 1.04$, the treatment effect of the baseline subgroup (i.e., intercept) is significant.

6 Discussion

This Chapter proposes a new data-driven method for studying treatment effect heterogeneity that notably improves interpretability of the heterogeneity and provides helpful guidance about subgroups with heterogeneous effects in terms of decision rules. Moreover, the proposed CRE methodology accommodates for well-known shortcomings of binary trees by providing a more stable, flexible and robust methodology to discover and estimate heterogeneous effects. Indeed, CRE is stable to sample-to-sample variations, leading to more reproducible results and, its flexibility allows for the discovery of a wider set of causal rules. Also, CRE provides robust results for the detection of causal rules in the presence of overlap between confounders and effect modifiers.

Though the CRE method makes inference using a smaller size of the sample due to sample-splitting, it maintains estimation precision at a similar level as other existing methods while providing an interpretable form of the treatment effect heterogeneity. The CRE method is a generic method that is completely compatible with existing methods for estimating heterogeneous treatment effect. The performance of CRE may vary with respect to the choice of existing methods to generate base decision rules in the discovery sample and intermediate values τ_i^* in the inference sample. Therefore, the CRE method can be thought of as a *refinement* process of the outputs produced by existing methods. If an estimation method for the CATE has great precision, then it is highly likely that it detects the treatment effect heterogeneity during the estimation procedure. When the CRE method is accompanied with this estimation method, the CRE method discovers the underlying treatment effect structure with high probability and represents this structure in an easy-to-interpret form. Indeed, a few simple rules are utterly important for public policy implications. However, when it comes to precision medicine, discovering a possibly lengthy rule that is specific to a patient could be of interest.

The proposed CRE method requires a researcher to specify some of the so-called tuning parameters. In particular, one should choose a number of trees that are generated for extracting decision rules, and the cutoff threshold in stability selection during the discovery step. Previous studies show that the performance is not much affected by specification of the parameters. Also, the studies provide a general guidance of how to specify the parameters. However, the optimal choice of the splitting ratio between the discovery and inference samples is not known yet. Even though the ratio (25%, 75%) is shown to have the best performance through simulation studies, we do not know whether this ratio works best for all real-world datasets. It may be possible to require a larger proportion of the discovery sample if data is sparse. On the contrary, a smaller proportion is needed when the underlying effect structure is simple.

Regardless of the splitting ratio, it is more important to choose a proper number of decision rules during the discovery step. The choice may depend on the questions that practitioners want to answer. For example, public policy makers generally want to discover a short list of risk factors. A few important subgroups defined by the risk factors are usually easy-to-understand, and further foster focused discussions about the assessments of potential risks and benefits of policy actions. Also, due to the restriction of resources, public health can be promoted efficiently when prioritized subgroups are available. Instead, a comparatively larger set of decision rules can be chosen, for instance, in precision medicine. An important goal is to identify patient subgroups that respond to treatment at a much higher (or lower) rate than the average (Loh et al., 2019). Also, identifying a subgroup that must avoid the treatment due to its excessive side effects can be valuable information. However, discovering only a few subgroups is likely to miss this extreme subgroup.

A number of extensions of the CRE method can be possible. First, the CRE method maintain the benefits of some of existing methods, i.e., asymptotic normality and unbiasedness. If an existing method can produce unbiased point estimates for $\tau(x)$ with valid confidence intervals, the CRE method can produce unbiased estimates for $\boldsymbol{\beta}$ with valid confidence intervals. Bayesian methods such as BART or BCF can be also used, and it is empirically shown that they perform really well. However, the validity of Bayesian inference such as constructing credible intervals remains as a future research question. Second, the discovery step of the CRE method can be considered as a dimension reduction procedure. We used a set of decision rules, but it may be possible to characterize the treatment effect heterogeneity using different approaches that are easy-to-interpret. Furthermore, one of the main issues that prevents evaluation of subgroup effects and heterogeneity in practice is small sample size, resulting in very low statistical power. Hence, it would be of interest, for further lines of research, to develop such an analysis in the context of an interpretable methodology, as the proposed CRE algorithm. Finally, we proposed an approach for sensitivity analysis of unmeasured confounding bias based on the inverse probability of treatment weighting estimator. A general approach for sensitivity analysis that can be compatible with a larger class of estimation methods would be helpful. Future research is needed for developing such sensitivity analysis.

Supplementary material for Chapter 2

7 Proofs

7.1 Proof of Theorem 1

By multiplying $(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T$ on the both sides of model (2.8), we have $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\nu}$. From the assumptions, $\text{plim}\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i = \mathbf{Q}$ and $\text{plim}\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i^T\nu_i = \text{plim}\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i^T(\epsilon_i + u_i) = 0$. By Slutsky's theorem, $\text{plim}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

7.2 Proof of Theorem 2

To prove normality, we use the expression $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\nu}$. By rewriting this, we have $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i)^{-1} \times N^{-1/2}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i^T\nu_i$. From the assumptions, we also have $\mathbb{E}(\tilde{\mathbf{X}}_i^T\nu_i) = 0$ and $\text{var}(\tilde{\mathbf{X}}_i^T\nu_i) = \mathbb{E}(\nu_i^2\tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i) < \infty$. Then, by the central limit theorem, $N^{-1/2}\sum_{i=1}^{N}\tilde{\mathbf{X}}_i^T\nu_i$ converges in distribution to $N(0, \Omega)$. By Slutsky's theorem and Cramer-Wold theorem, $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to $N(0, \mathbf{Q}^{-1}\Omega\mathbf{Q}^{-1})$.

8 Comparison between BCF, BART, IPW, OR for rules' discovery

As we discussed in Section 2, one of the attractive features of the CRE methods is that we can consider many approaches to estimate the individual treatment effect τ_i. Among many existing methods, one approach to note is the inverse probability weighting estimator:

$$\hat{\tau}_i^{IPW} = \left(\frac{W_i}{\hat{e}(\mathbf{X}_i)} - \frac{1 - W_i}{1 - \hat{e}(\mathbf{X}_i)} \right) Y_i \tag{2.12}$$

where $\hat{e}(\mathbf{X}_i)$ is the estimate of the propensity score $e(x)$ at $\mathbf{X}_i$ and IPW denotes inverse probability weighting. The estimate $\hat{e}(\mathbf{X}_i)$ can be obtained by fitting a logistic regression on $(W_i, \mathbf{X}_i)$. The estimator is validated based on the identification result (2.3). Although $\hat{\tau}_i^{IPW}$ is an unbiased estimator of $\tau(x)$ (i.e., $\mathbb{E}[\hat{\tau}_i^{IPW} | X = x] = \tau(x)$), the transformed value $\hat{\tau}_i^{IPW}$ can be highly fluctuating when $\hat{e}(\mathbf{X}_i)$ is close to 0 or 1. To avoid extreme values of $\hat{\tau}_i^{IPW}$, we can instead use the stabilized version $\hat{\tau}_i^{SIPW}$ (Hirano et al., 2003),

$$\hat{\tau}_i^{SIPW} = \left\{ \left(\frac{1}{N} \sum_{i=1}^{N} \frac{W_i}{\hat{e}(\mathbf{X}_i)} \right)^{-1} \frac{W_i}{\hat{e}(\mathbf{X}_i)} - \left(\frac{1}{N} \sum_{i=1}^{N} \frac{1 - W_i}{1 - \hat{e}(\mathbf{X}_i)} \right)^{-1} \frac{1 - W_i}{1 - \hat{e}(\mathbf{X}_i)} \right\} Y_i. \tag{2.13}$$

Although $\hat{\tau}_i^{IPW}$ or $\hat{\tau}_i^{SIPW}$ is enough to use as an estimate of τ_i, another approach of imputing missing potential outcomes can be considered. Without estimating the propensity score, functions for two potential outcomes, say $m_1(x) = \mathbb{E}[Y | W = 1, \mathbf{X} = x]$ and $m_0(x) = \mathbb{E}[Y | W = 0, \mathbf{X} = x]$, are estimated. Then missing potential outcomes are imputed by the estimated functions $\hat{m}_0$ and $\hat{m}_1$. For instance, if $W_i = 0$, $Y_i(0)$ is observed as Y_i and $Y_i(1)$ is imputed by $\hat{m}_1(\mathbf{X}_i)$. The unit level treatment effect can be estimated by either subtracting the observed outcome and imputed counterfactual, i.e. $\tau_i^{OR} = Y_i^{obs} - (\hat{m}_1(\mathbf{X}_i) \cdot (1 - W_i) + \hat{m}_0(\mathbf{X}_i) \cdot W_i)$, or by subtracting the imputed potential outcomes, i.e., $\tau_i^{BART} = \hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)$. We refer to the former methodology as outcome regression (OR) and to the latter as BART imputation (Hill, 2011). Indeed, we implement both these methods for the estimation of the unit level treatment effect using the Bayesian Additive Regression Tree (BART) algorithm Chipman et al., 2010.

Here, we show that the BCF method for the estimation of τ_i has the better performance as compared to other approaches such as inverse probability weighting, outcome regression (OR) and BART (Chipman et al., 2010; Hill, 2011). In order to evaluate the ability of discovering decision rules, two factors are considered in line with similar simulations scenarios (Bargagli-Stoffi et al., 2019): (1) how many rules are discovered and (2) among the discovered rules, how many times true rules are captured. We implement the simulation scenario introduced in Subsection 2.2 with uncorrelated covariates, linear confounding and 2,000 data points. The obtained results are depicted in Figure 16. The four different plots in the Figure show the variation in the number of correct rules (first row) and the number of detected rules (second row) as the effect size increases. The plots in the first column depict the results in the case of two true causal rules, while in the second column depict the results in the case of four true causal rules.

In the case of two true causal rules, BCF, BART and OR perform in a very similar manner with respect to their ability to identify the true causal rules, while in the case of four true causal rules there is a clear advantage in using BART or BCF over OR. IPW is consistently outperformed in both the scenarios. With respect to the number of rules detected, OR and IPW are the more "conservative" methods (i.e., smaller number of detected rules) while BART is the less "conservative" method. These results inform our choice to implement the CRE-BCF method, as BCF shows the best performance in terms of correctly detected rules, while it detects consistently less rules than the other best performing methodology (BART) and it provide higher computational scalability (i.e., we need to estimate just one model and not two as in the case of BART).

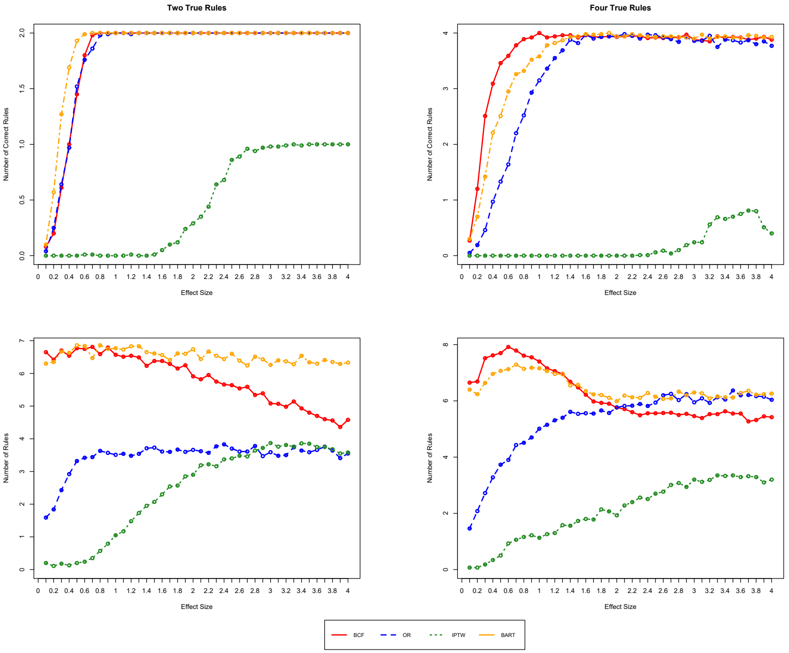


Figure 16: Comparison between BCF, BART, IPW, OR for rules' discovery. The plots show the variation in the number of correct rules (first row) and the number of detected rules (second row) as the effect size increases, in the cases of two true rules (first column) and four true rules (second column).

9 Detailed simulation results

Table 14 depicts the performance of CRE-BCF for the simulation scenario introduced in Subsection 4.1 with two and four causal rules and with 1,000 and 2,000 data points. We consider 100 simulated datasets for each effect size and we provide both the average number of Correctly Discovered Rules (CDR), the proportion of times when the correct rules are discovered (π) and the average number of Discovered Rules (DR). As the effect size k increases CRE is almost always able to spot the true causal rules. Even for smaller effect sizes, where the causal rules are not statistically significant (i.e., when the effect size is smaller than one), the performance of CRE is excellent both in the case with two and four true rules.

Table 14: Performance of the CRE-BCF method in discovering the true underlying causal rules

Linear Scenario												
	Two Rules						Four Rules					
	1,000			2,000			1,000			2,000		
k	CDR	π	DR	CDR	π	DR	CDR	π	DR	CDR	π	DR
0.1	0.03	0.00	6.53	0.04	0.01	6.41	0.15	0.01	6.54	0.25	0.01	6.57
0.2	0.05	0.01	6.54	0.22	0.05	6.73	0.33	0.04	6.51	1.06	0.14	6.88
0.3	0.13	0.02	6.50	0.55	0.15	6.57	0.92	0.09	6.66	2.18	0.35	6.82
0.4	0.34	0.09	6.54	0.92	0.33	6.60	1.86	0.29	7.04	2.96	0.48	7.55
0.5	0.64	0.17	6.64	1.43	0.61	6.82	2.67	0.40	7.44	3.62	0.83	7.68
0.6	1.05	0.42	6.59	1.86	0.90	6.80	3.03	0.56	7.48	3.76	0.83	7.66
0.7	1.37	0.59	6.84	1.99	0.99	6.67	3.48	0.71	7.59	3.84	0.89	7.60
0.8	1.89	0.90	6.61	2.00	1.00	6.65	3.64	0.79	7.61	3.92	0.94	7.51
0.9	1.96	0.96	6.77	2.00	1.00	6.66	3.71	0.82	7.66	3.94	0.95	7.39
1.0	1.99	0.99	6.68	2.00	1.00	6.55	3.93	0.96	7.71	3.97	0.97	7.35
1.1	1.98	0.98	6.79	2.00	1.00	6.51	3.90	0.92	7.51	3.97	0.98	7.24
1.2	1.99	0.99	6.63	2.00	1.00	6.36	3.93	0.95	7.62	3.94	0.96	6.88
1.3	2.00	1.00	6.65	2.00	1.00	6.67	3.97	0.97	7.64	3.94	0.94	6.68
1.4	2.00	1.00	6.76	2.00	1.00	6.30	3.98	0.98	7.59	3.96	0.96	6.82
1.5	2.00	1.00	6.65	2.00	1.00	6.31	3.92	0.95	7.38	3.88	0.89	6.52
1.6	2.00	1.00	6.65	2.00	1.00	6.29	3.95	0.95	7.18	3.94	0.94	6.23
1.7	2.00	1.00	6.61	2.00	1.00	6.26	3.93	0.93	7.16	3.93	0.93	6.04
1.8	2.00	1.00	6.47	2.00	1.00	6.01	3.90	0.90	7.16	3.96	0.96	6.01
1.9	2.00	1.00	6.50	2.00	1.00	6.03	3.95	0.95	6.94	3.85	0.85	5.84
2.0	2.00	1.00	6.24	2.00	1.00	5.93	3.93	0.93	6.90	3.96	0.97	5.84

10 Additional simulations

In this Section of the Appendix, we run an additional set of simulations introducing variations in the data generating process described in Section 4.1. In particular, we introduce correlation between each covariate in the $\mathbf{X}$ matrix (i.e., all the covariates have a 0.3 correlation with each other). Such correlations were introduced to investigate whether or not correlated covariates negatively affect the ability of CRE-BCF to discover the true causal rules. Indeed, it could be possible that CRE-BCF faces harder times in correctly picking the variables that are responsible for the heterogeneous effects, as all the variables are correlated with each other. Figure 17 depicts the results for this new simulation setting in the case of 1000 data points and 2 and 4 true causal rules, respectively. No variation was found in the performance of the model with respect to the original data generating process with no correlation in the covariate matrix. In fact, the proposed model shows a very good simulated performance in correctly detecting the causal rules both with and without correlation between the variables.

Moreover, we generate non-linearities in the confounding $f(\mathbf{X})$. In particular, the data generating process introduced in Section 2.2 is reworked as follows: $y = y_0 \cdot (1 - z_i) + y_1 \cdot z_i + exp\{X_1 - X_2 \cdot X_3\}$. This non-linear $f(\mathbf{X})$ is introduced in order to check the robustness of the CRE-BCF model to non-linear confounding. Indeed, we can argue that in many real-world applications the confounders can interact with each other and can have non-linear associations with the output. Again, we generate two different scenarios with 2 and 4 true causal rules and 1000 data points. From Figure 18 we find that this kind of non-linear confounding is not harmful for our model, but it even slightly increases its ability of correctly discover the true causal rules. This is due to the fact that both BCF (**hahn2020bayesian**) and CRE are able to deal with non-linearities in a excellent way.

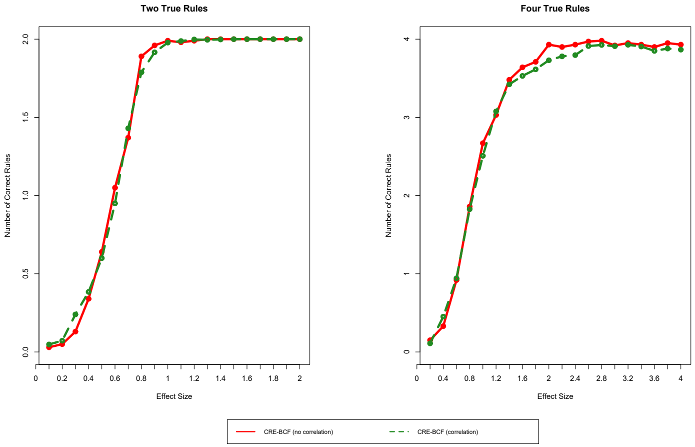


Figure 17: Average number of correctly discovered rules for CRE-BCF in the case of two true rules (left panel) and four true rules (right panel).

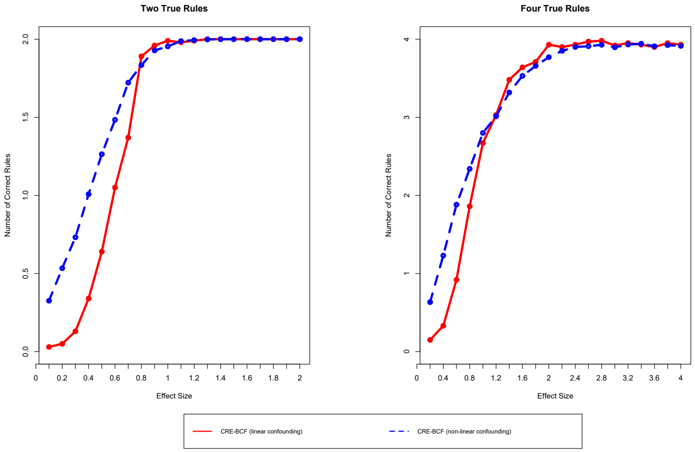


Figure 18: Average number of correctly discovered rules for CRE-BCF in the case of two true rules (left panel) and four true rules (right panel).

Part 2
Machine learning for prediction

Chapter 3

Assessing sensitivity of machine learning predictions

Machine learning is rapidly growing in popularity in social sciences (Athey & Imbens, 2019). As we highlighted in the Introduction of the present Dissertation, this growth is due to the fact that machine learning provides an excellent set of tools to address a large number of issues regarding prediction, causal inference, theory testing and development of new data sources (Mullainathan & Spiess, 2017). The recent rise in the usage of machine learning algorithms across various scientific disciplines has been mainly driven by the staggering performance of these algorithms in predictive tasks. However, the usage of off-the-shelf machine learning methodologies to perform predictions should be done in a careful and thoughtful way. Today, little has been done to explore how sensitive machine learning predictions are to potentially unobserved predictors.

On the methodological side, this Chapter accounts and accommodates for this shortcoming by proposing a general algorithm that assesses how the omission of an unobserved predictor could affect the predic-

This Chapter is based on Bargagli-Stoffi, De Beckker, De Witte and Maldonado (2020).

tions and the performance of the model. As such, we obtain estimates for, what we call, *sensitivity of prediction analysis*. The algorithm that we propose is general enough to be applied to both Bayesian methodologies (i.e., Bayesian Additive Regression Trees [Chipman, 2010]) and frequentist techniques (i.e., Random Forests [Breiman, 2001]). In particular, the method that we propose aims at answering policy-relevant questions such as: *"is the machine learning model getting enough signal from the data?"* and *"how much would an unobserved predictor impact the model's predictions and its performance?"*. To do so, we generate a synthetic predictor with a high explanatory power (i.e., high correlation with the outcome) but that is uncorrelated with all the predictors in the model. Then, we check how the inclusion of this additional variable in the machine learning model modifies its predictions and accuracy. We assume that the model at hand is able to catch all the *signal* in the data if its forecasts and its predictive ability are not extensively modified by the inclusion of the synthetic predictor. In order to test this, we check if: (i) the unit level predictions of the model with the synthetic predictor (*augmented model*) fall within the confidence interval of the unit level prediction of the model that we are testing; and (ii) the predictive performance of the augmented model is statistically different from the performance of the original model. The approach that we develop here is applied in the case of a continuous outcome variable, but it can be easily extended to a binary outcome variable. Moreover, we make this methodology robust by introducing, in a second step, a set of correlations between the synthetic predictor and the most important predictors in the original model. This second case more closely mimic the case of real-world applications where predictors are correlated within each other.

On the applied side, we innovate the literature by introducing the first application of supervised machine learning in the financial literacy literature. First, we use machine learning to predict financial literacy scores (FLS) for students in a region of Belgium where these scores are not observed. We train our model using the PISA data (OECD, 2017b) for the Flemish region of Belgium were the FLS are observed. Then, we perform the prediction for the missing FLS of students in the Wallonia region of

Belgium. We use an extensive set of student and school-level variables as predictors in our machine learning model. Second, we use machine learning to identify the characteristics of students with outlying predicted test scores. In our case, the main focus is on students with lower predicted FLS. Indeed, machine learning can help researchers and policymakers to understand the characteristics of students who have low predictions for financial literacy. These insights enable policy-makers to target interventions to lower performing students in order to improve their performances. Third, we employ the novel sensitivity analysis to assess the robustness of our analyses. Sensitivity is central in the case of our application, as we deal with sensitive data such as students' scores in a standardized assessment on financial literacy. In particular, it is deemed important to assess how sensitive the predictions of student financial literacy scores are to a potentially unobserved predictor with high explanatory power.

Our main results can be summarized as follows. The performance of the proposed model is good as we reach an estimated adjusted R^2 of roughly 73% (this performance measure results from a 10-folds cross-validation[1]). Moreover, the chosen machine learning technique, Bayesian Additive Regression Trees (Chipman et al., 2010), allows us to get draws from the posterior distribution of the predicted observations and construct credible intervals for the unit level predictions. Hence, we are able to detect the predictions that fall outside the 95% credible intervals for the mean posterior predicted values. We use this information to identify a set of outlier predictions. On one side, this analysis shows that the predictive probability of having a low financial literacy scores is the largest for students with lower scores for reading and math in the PISA test. On the other side, the students' background plays a critical role: students with the largest predicted low financial literacy scores are often individuals from families where the school language is not spoken at home and the parents have a poor educational background. Moreover, results from the sensitivity of predictions analysis show that the inclusion of an un-

[1] 10-folds cross-validation is a well known way to assess the performance of a predictive model in machine learning.

observed predictor would not affect the predictions and the predictive ability of the model in a sizable manner. This hints at the fact that our model is already getting enough *signal* to perform its predictions.

We argue that the proposed methodologies have applications beyond the financial literacy literature. Indeed, in applied sciences, there is a need to introduce uncertainty of predictions as machine learning is more and more extensively used to perform predictions on very sensitive manners: i.e., patients' life expectancy (Kleinberg et al., 2015), human capital selection (Chalfin et al., 2016), unemployment insurance effectiveness (Chernozhukov et al., 2017), students' learning abilities (Delen, 2010; S. B. Kotsiantis, 2012; S. Kotsiantis et al., 2004), and so on. In this spirit, the proposed methodology and the motivating application introduced in this paper are a first step towards a more transparent way to account for predictions stability and uncertainty in machine learning applications in social sciences.

The remainder of the Chapter is organized as follows. In Section 1 we introduce the machine learning methodologies used and developed in this Chapter; Section 3 describes the PISA data used to illustrate the proposed methodology; in Section 3 we discuss the results obtained from the application on the PISA data; and Section 4 concludes the Chapter with a discussion of the methods and results and with new potential lines of research[2].

1 Methodology

This Section discusses in detail the machine learning technique used for prediction (Subsection 1.1) and its extension in order to detect observations with outlying values which paves the way to targeted policies. In Subsection 1.2, we introduce the novel methodologies for sensitivity analysis that extend the usage of machine learning algorithms from pointwise predictions to robustness analysis regarding these predictions.

[2] The `R` and `Stata` codes used for the analysis, together with the data and the functions for the machine learning analysis are publicly available on the GitHub page of the corresponding author (`https://github.com/fbargaglistoffi`).

1.1 Machine learning for predictions and targeted policies

In order to make predictions, we have to train a machine learning algorithm. As underlying method, we rely on BART (Chipman et al., 2010), which is a Bayesian ensemble of trees method[3]. BART was shown to have an excellent performance in both prediction tasks (Bargagli-Stoffi, Riccaboni, et al., 2020; Hernández et al., 2018; Linero, 2018; Linero & Yang, 2018; Murray, 2017; Starling et al., 2018) and in causal inference tasks (Bargagli-Stoffi et al., 2019; P. R. Hahn et al., 2020; Hill, 2011; K. Lee et al., 2020; Logan et al., 2019; Nethery et al., 2019). This is due to the particular flexibility of this method and the fact that it overcomes potential issues connected with other machine learning methodologies such as the Classification And Regression Tree (CART) algorithm (Breiman et al., 1984) and the Random Forest (RF) algorithm (Breiman, 2001). Moreover, BART allows researchers to get draws from the posterior distribution of the predicted values. In Bayesian statistics, the posterior predictive distribution indicates the distribution of predicted data based on the data one has already seen (Gelman et al., 2014). Hence, the posterior predictive distribution can be used to predict new data values and to draw inference around the distribution of these predicted values.

BART can be considered a “refined” version of the RF methodology. Both these methodologies found their foundation in the CART algorithm. CART is a widely used algorithm for the construction of trees where each node is split into only two branches (i.e., binary trees). Breiman (2001) has shown that the accuracy of the predictions of CART can be dramatically improved by averaging the prediction of an ensemble of uncorrelated trees. For additional details on how CART and RF work, we refer the reader to Chapter 1 where these algorithms were introduced in detail, and to the seminal papers by Breiman et al. (1984) and Breiman (2001).

As we saw in Chapter 1, similarly to the RF algorithm, BART is a sum-of-trees ensemble methodology. However, BART relies on a fully

[3] Extension to other machine learning algorithms such as random forest (Breiman, 2001) is straightforward.

Bayesian probability model to obtain its predictions (Kapelner & Bleich, 2016). As we are in a predictive settings, let us define with Y the outcome vector, with y_i the outcome for a generic unit i, and with $\mathbf{X}$ the $N \times P$ matrix of predictors[4] (where P is the number of predictors), with X_i the P-dimensional vector of predictors for i (with $i = 1, ..., N$). Then, using a more compact notation than the one used in Chapter 1, we can express BART as:

$$Y = \sum_{j=1}^{J} \mathcal{T}_j(\mathbf{X}; \mathcal{D}_j, \mathcal{M}_j) + \epsilon. \tag{3.1}$$

where the J distinct binary trees are denoted by $\mathcal{T}(\mathbf{X}; \mathcal{D}_j, \mathcal{M}_j)$. $\mathcal{T}$ is a function that sorts each unit into one of the sets of m_j terminal nodes, associated with mean parameters $\mathcal{M}_j = \{\mu_1, ..., \mu_{m_j}\}$, based on a set of decision rules, $\mathcal{D}_j$. ϵ is an error term and is typically assumed to be independent, and identically normally distributed when the outcome is continuous (Chipman et al., 2010). For a detailed discussion on the set of priors used in BART the reader is referred to Chapter 1 and the seminal paper by Chipman et al. (2010). Here, we just want to highlight that the aim of these priors is to "regularize" the algorithm, which prevents single trees to dominate the overall fit of the model (Kapelner & Bleich, 2016).

BART allows the researcher to tune the variables' importance by departing from the original formulation of the Random Forest algorithm where each variable is equally likely to be chosen from a discrete uniform distribution (i.e., with probability $\frac{1}{p}$) to build a single tree learner. These Bayesian tools give researchers the possibility to mitigate the overfitting problem of RFs and to tune the algorithm with prior knowledge. Moreover, BART has shown a consistently strong performance under "default" model specifications. This is a highly valuable characteristic of BART as it reduces its dependence on the choice of parameters done by the researcher as well as the computational time and costs related to cross-validation (Nethery et al., 2019).

To see in more detail how BART works, let us introduce a simple case in which we have a study sample ($\Omega = \{\mathbf{X}, Y\}$) for which we observe

[4] In the machine learning literature, the variables that are used to fit the model are usually referred to as *predictors*.

the set of predictors $\mathbf{X}$, but the vector of outcomes is observed just for a subsample of the study population, Y^{obs}. Hence, we can indicate with Y^{mis} the vector of missing outcomes in the study sample, where $Y = Y^{mis} \cup Y^{obs}$. To perform the imputation, BART collects a sample from the posterior distribution of $\theta = \{\sigma^2, \mathcal{D}_j, \mathcal{M}_j\}$, $p(\theta|Y^{obs})$ using the Bayesian backfitting algorithm proposed by Chipman et al. (2010). An imputed value for $\hat{Y}^{mis}$ can be obtained by sampling from the posterior predictive distribution (ppd):

$$p(Y^{mis}|Y^{obs}) = \int p(Y^{mis}|Y^{obs}, \theta) \cdot p(\theta|Y^{obs}) \, d\theta. \tag{3.2}$$

From equation (3.2) we obtain a vector of predicted values for the missing outcome observations. Then, we can use standard techniques employed for outlier detection, such as the ones proposed by J. Miller (1991) (i.e., two/three standard deviation from the mean of the distribution) and Leys et al. (2013) (i.e., two/three absolute deviations from the median), to detect the predictions with values that are consistently further away from the mean or the median of the predicted distribution, respectively. The nice feature of such techniques, is that they can be manually tuned in order to include units with more or less extreme predicted values. This type of analyses can therefore be highly relevant for policy-makers to spot observations with more or less extreme predicted outcome values. Indeed, it can be the case that policy-makers could be interested in targeting their intervention to a specific subpopulation, based on the values of an unobserved outcome (e.g., provide additional learning material to more vulnerable students in a region where financial literacy is not observed). Moreover, information on the factors related to these "high or low levels" of the predicted outcome are not only useful to identify potential subgroups but also to reveal which factors are associated with specific levels of the outcome.

1.2 Sensitivity of predictions analysis

The ability of machine learning techniques to provide accurate and *robust* predictions is key. As predictions always hold a degree of uncertainty, we argue that there is a need for specific techniques that enable researchers to assess their *sensitivity*. This is particularly true as the usage of machine learning for prediction policy problems has been increasing over the last few years (Kleinberg et al., 2015; Mullainathan & Spiess, 2017).

The tasks of the *sensitivity of predictions analysis* introduced here is to assess how the omission of an unobserved predictor could affect both the forecasts of the model and its performance. Indeed, we argue that a desirable property of any predictive model is to be *robust* to potentially unobserved predictors. In a predictive scenario, we can define *robustness* as the model's ability to provide stable predictions and a stable performance over the inclusion of additional variables in the model. Robustness of predictions is a desirable property as volatile predictions may be completely overturned by the inclusion of an additional predictor in the model. Stability of predictions per-se is not necessarily a desideratum: a model that is completely unable to properly predict a certain outcome may be stable but highly inaccurate. We argue that the stability of the model depends on the quality of the additional predictor included. Indeed, as we include a novel variable with a high explanatory power in a poorly performing model (i.e., a model with highly inaccurate predictions), the model will show a boost in its predictive power driven by an increase in the precision of its predicted values. Hence, it is intuitive to see how such a sensitivity of prediction analysis is important for at least two reasons. First, it can show how much an additional predictor could boost the performance of the model, indicating how much the model can be improved by taking into account additional, potentially unobserved predictors. Second, it reveals the variability in the predictions induced by novel predictors: the more stable they are, given a high predictive performance, the more the model can be trusted for predictive policy tasks.

Such a sensitivity of prediction analysis is directly inspired by its causal inference sensitivity analysis counterpart (Ichino et al., 2008; Rosenbaum & Rubin, 1983a). In particular, Ichino et al. (2008) propose a simulation-based sensitivity analysis to assess the robustness of causal estimands to failures in the unconfoundedness assumption (i.e., independence between the treatment and the potential outcomes). The authors build on the works by Rosenbaum and Rubin (1983b) and Rosenbaum and Rubin (1983a) and simulate an unobserved confounder that is both associated with the potential outcome and the treatment. Then, they introduce this additional confounder in the model for the estimation of the Average Treatment effect on the Treated (ATT). They repeat the estimation procedure many times using a different set of values for the simulated confounder and compare these estimates with the original estimate of the ATT obtained without the inclusion of the simulated confounder. Such comparison informs scholars on how robust the original causal estimands are to different configurations used for the construction of the confounder (and, in turn, different levels of violations of the unconfoundedness assumption).

While using a similar intuition, the sensitivity analysis that we implement here is fundamentally different because we deal with a prediction setting rather than a causal inference one. In fact, as we deal with prediction problems, we depart from the conceptual meaning of sensitivity analysis, in causal inference, as a strategy to assess the robustness of an estimators to an underlying and untestable set of assumptions. In this sense, the sensitivity analysis that we introduce in this Chapter is not aimed at testing any model assumption or identification assumption. Conversely, our aim is to provide a direct assessment of the quality of the model and its predictions. In particular, the added value of the proposed approach is in those situations where one has a well-performing predictive model and wants to assess whether or not it's worthwhile to continue the search for new predictors.

The sensitivity of prediction analysis is performed by generating a synthetic predictor and checking if (and how) its inclusion in the set of predictors changes the model's predictions and, in turn, its performance. Let us define with R a P-dimensional vector for the synthetic predictor. As we want this predictor to have a good explanatory power with respect to the outcome vector Y. Hence, we will design it it to have a high correlation between the outcome and the synthetic predictor but to be uncorrelated with all the predictors in the model.

Before going into the details of how to construct the synthetic variable, let us discuss the implications of constructing it in such a way. While a high correlation with the outcome and a zero correlation with the observed predictors guarantees a high explanatory power (i.e., a large share of the variation in the outcome being explained by the synthetic variable, controlling for the other predictors), it does not guarantee that the synthetic variable is also the one with the highest predictive power. Indeed, it may be the case that in a non-linear setting, the interaction between two variables with lower explanatory power and the outcome could carry more information than a predictor with a higher explanatory power than these two individual variables. However, such interaction patterns between the predictors and the outcome are hidden in the data (in some way, the ultimate goal of machine learning methodologies is to discover such underlying patterns), and is not possible to reproduce them in any meaningful way. Hence, constructing a variable with a high predictive power, may not lead to the best absolute model performance but it is (i) feasible (while it may not be feasible to construct a variable with the highest predictive power due to the complex, non-linear patterns of interactions between the predictors and the outcome) and; (ii) meaningful as the explanatory power of a variable is relevant also in a prediction setting (especially when the ultimate users are the policy-makers).

Let us show in detail how to compute the synthetic predictor. We start from sampling N independent realizations from a normally distributed continuous variable: $\tilde{R} \sim \mathcal{N}(0, 1)$[5] . We define $\tilde{\mathbf{R}}$ the $N \times (P + 2)$ matrix that stacks the output vector Y, $\tilde{R}$ and the matrix of observed variables $\mathbf{X}$ such that:

$$\tilde{\mathbf{R}}_{(N\times(P+2))} = \begin{bmatrix} Y_{(N\times1)} & \tilde{R}_{(N\times1)} & \mathbf{X}_{(N\times P)} \end{bmatrix}. \tag{3.3}$$

We define with $\hat{\mathbf{K}}$ the estimated correlation matrix of $\tilde{\mathbf{R}}$:

$$\hat{\mathbf{K}}_{((P+2)\times(P+2))} = \begin{pmatrix} 1 & \hat{\rho}(Y, \tilde{R}) & \dots & \hat{\rho}(Y, X_p) \\ \hat{\rho}(\tilde{R}, Y) & 1 & \dots & \hat{\rho}(\tilde{R}, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}(X_p, Y) & \hat{\rho}(X_p, \tilde{R}) & \dots & 1 \end{pmatrix} \tag{3.4}$$

[5] The distribution of this variable mimics the distribution of the synthetic predictor and can be directly changed by the researcher.

and with $\tilde{\boldsymbol{\Sigma}} = diag(\hat{\sigma}^2_{\tilde{Y}}, \hat{\sigma}^2_{\tilde{R}}, \hat{\sigma}^2_{X_1} ..., \hat{\sigma}^2_{X_p})$ the $(P+2) \times (P+2)$ diagonal matrix of estimated variances, where $\hat{\sigma}^2_{\tilde{Y}}$ is the estimated variance of the output, $\hat{\sigma}^2_{\tilde{R}}$ is the estimated variance of $\tilde{R}$ and $\hat{\sigma}^2_{\tilde{X}_j}$ is the estimated variance of the j-th variable in $\mathbf{X}$. Moreover, if we define with $\mathbf{C}$ the symmetric, idempotent centering $N \times N$ matrix:

$$\mathbf{C} = \mathbf{I} - \frac{1}{N}\mathbf{1} \tag{3.5}$$

where $\mathbf{1}$ is a $N \times N$ matrix of 1s, we can center and scale the $\tilde{\mathbf{R}}$ matrix as follows:

$$\dot{\mathbf{R}} = \mathbf{C}\tilde{\mathbf{R}}\tilde{\boldsymbol{\Sigma}}^{-1} \tag{3.6}$$

where $\dot{\mathbf{R}}$ is a multivariate normal matrix. This is similar to the type of scaling used, for example, in standard normal variate correction (Barnes et al., 1989), and corresponds to a variation of the whitening (or sphering) transformation . Then, we introduce no-correlation within the columns in $\tilde{\mathbf{R}}$ as follows[6]:

$$\mathbf{L}^{-1}\dot{\mathbf{R}} \tag{3.7}$$

where $\mathbf{L}$ is a lower triangular matrix obtained through the Cholesky decomposition of $\tilde{\mathbf{K}}$:

$$\mathbf{L}\mathbf{L}^T = \tilde{\mathbf{K}}. \tag{3.8}$$

The matrix that we obtain in 3.7 is a set of realizations of uncorrelated multivariate normal random variables. At this point, we can reintroduce the correlation between each column:

$$\mathbf{U}\mathbf{L}^{-1}\dot{\mathbf{R}} \tag{3.9}$$

where $\mathbf{L}$ is obtained thought as the Cholesky decomposition of $\mathbf{K}$:

$$\mathbf{U}\mathbf{U}^T = \mathbf{K} \tag{3.10}$$

and $\mathbf{K}$ is a newly defined correlation matrix:

$$\mathbf{K}_{((P+2)\times(P+2))} = \begin{pmatrix} 1 & \widetilde{\rho}(Y,R) & \dots & \hat{\rho}(Y,X_p) \\ \widetilde{\rho}(R,Y) & 1 & \dots & \widetilde{\rho}(R,X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}(X_p,Y) & \widetilde{\rho}(X_p,R) & \dots & 1 \end{pmatrix} \tag{3.11}$$

[6] This is possible as far as $\mathbf{K}$ is a symmetric positive-definite matrix.

where $\widetilde{\rho}(Y, R)$ is the correlation between the synthetic predictor R and the outcome, that is set directly by the researcher, and $\widetilde{\rho}(X_j, R)$ is the correlation between the synthetic predictor the $j-$th regressor. These correlations are set to be equal to zero in our setting:

$$\mathbf{K}_{((P+2)\times(P+2))} = \begin{pmatrix} 1 & \widetilde{\rho}(Y,R) & \dots & \hat{\rho}(Y,X_p) \\ \widetilde{\rho}(R,Y) & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}(X_p,Y) & 0 & \dots & 1 \end{pmatrix} \tag{3.12}$$

In a further step, we make this algorithm robust by introducing correlations between the synthetic variable and the best predictor, to closely mimic the case of real-world applications where predictors are correlated within each other. The choice of the correlations pattern is left to the researcher, and can be informed by previous analyses on the relative importance and correlation of the predictors.

Finally, we scale the data back to their original distribution:

$$\mathbf{R} = \mathbf{C}^{-1}\mathbf{U}\mathbf{L}^{-1}\dot{\mathbf{R}}\tilde{\mathbf{\Sigma}} \tag{3.13}$$

and "extract" the newly generated synthetic predictor R from the $\mathbf{R}$ matrix. In this specific case, the synthetic predictor corresponds to the second column of the matrix $\mathbf{R}$. However, its relative position depends on how the matrix $\tilde{\mathbf{R}}$ is generated.

At this point, we can rework the model in (3.1) to include the set of stack predictors $\mathbf{G}$:

$$\sum_{l=1}^{L} \mathcal{T}_l(\mathbf{G}; \mathcal{D}_l, \mathcal{M}_l) + \psi \tag{3.14}$$

where $\mathbf{G}$ is:

$$\mathbf{G}_{(N\times(P+1))} = \big[R_{(N\times1)}\ \mathbf{X}_{(N\times P)}\big]. \tag{3.15}$$

Finally, we can compare the predictions of the new *augmented* model in (3.14) with the ones of the *original* BART model in (3.1). To do so, we run the two models on b bootstrapped samples $b = 1, ..., B$ to get a set of predicted values at each iteration

$$\sum_{j=1}^{J} \mathcal{T}_{j,b}(\mathbf{X}; \mathcal{D}_j, \mathcal{M}_j) + \epsilon; \tag{3.16}$$

$$\sum_{l=1}^{L} \mathcal{T}_{l,b}(\mathbf{G}; \mathcal{D}_l, \mathcal{M}_l) + \psi. \tag{3.17}$$

Once we get the unit level predictions for these models at each iteration step b ($\hat{y}_b(x_i)$ and $\hat{y}_b(g_i)$, respectively), we can run a series of comparisons between both the distribution of the predictions and the performance of augmented model as compared to the original model. To assess the statistical difference between the unit level predictions' distributions we test the difference between the mean of the unit level prediction of the augmented model, $\bar{\hat{y}}(x_i)$, and the mean of the unit level predictions of the original model, $\bar{\hat{y}}(r_i)$

$$\bar{\hat{y}}(x_i) = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_b(\mathbf{X} = x_i) \quad \text{and} \quad \bar{\hat{y}}(r_i) = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_b(\mathbf{G} = g_i). \tag{3.18}$$

Moreover, to assess whether or not there is an improvement in the performance of the augmented model, we run a series of tests on the difference of both the Root-Mean-Squared-Error of prediction and the adjusted R^2 of the two models[7].

These tests provide a measure of the extent to which the inclusion of a new regressor that is correlated with the outcome and uncorrelated with the observed predictors, would impact the model's predictions and performance. As the correlation coefficients between the synthetic predictor and the outcome are defined by the scholar herself[8], this impact can be estimated for different correlation settings. The more similar the unit level predictions, and the more similar the performance of the original and the augmented model, the more one can argue that the available signal is enough for stable and precise predictions.

The table Algorithm 3 summarizes the algorithm introduced in this Subsection and its main steps[9].

[7] For additional details on this set of tests in the online appendix.

[8] With the constraint of $\mathbf{K}$ being a symmetric positive-definite matrix.

[9] Here, we introduce our algorithm in its most general version. However, running B times the machine learning model at hand can be computationally intensive, especially for less scalable machine learning methodologies. In the case of BART, to reduce such computational burden, one can directly sample the unit level predictions from the posterior

Algorithm 3 Overview of the sensitivity analysis

The steps of the algorithm:

1. create a new matrix of predictors $\mathbf{G}$ stacking the observed predictors $\mathbf{X}$ and a synthetic predictor R correlated with the outcome Y and uncorrelated to all the variables in $\mathbf{X}$;
2. generate two models, one including only the observed predictors $\mathbf{X}$ (*original model*) and one including the observed predictors and the synthetic predictor $\mathbf{G}$ (*augmented model*);
3. run the two models on b bootstrapped samples ($b = 1, ..., B$) to get a set of predicted values at each iteration:

$$\hat{Y}_b(\mathbf{X}) = \sum_{j=1}^{J} \mathcal{T}_{j,b}(\mathbf{X}; \mathcal{D}_j, \mathcal{M}_j) + \epsilon \tag{3.19}$$

$$\hat{Y}_b(\mathbf{G}) = \sum_{l=1}^{L} \mathcal{T}_{l,b}(\mathbf{G}; \mathcal{D}_l, \mathcal{M}_l) + \psi; \tag{3.20}$$

4. run a series of tests comparing the original and the augmented model with respect to:
   (a) their unit level predictions;
   (b) their performance.

2 Data

To illustrate the proposed methodologies for prediction, subgroups' targeting and sensitivity, we apply the methodology to the 2015 data of the OECD's Program for International Student Assessment (PISA) for Belgium. The structure of the Belgian PISA 2015 data offers an interesting example of forecasting missing scores on a standardised student assess-

predictive distribution. For the RF algorithm, we argue that researchers can invoke the statistical properties introduced by Wager and Athey (2018) to construct reliable confidence intervals and statistical tests for each unit level prediction.

ment. In particular, in the special case of Belgium, all regions of the country participated in the general assessment of PISA, while only selected regions participated in the financial literacy assessment[10] This structure of the data allows us to use a common set of predictors for all regions from the general assessment in order to predict financial literacy outcomes for students in the regions that did not participate in this part of PISA.[11].

We choose the Belgian case to illustrate the forecasting methodology, since the regions of the small country have many similarities and common policies that justify the assumption of a common relationship between the predictors and the outcome variable. At the same time, the Belgian regions differ in population characteristics as well as education policies which suggests that the missing scores in Wallonia are unlikely to be identical to those in Flanders. Indeed, as shown in Table 20 in Appendix 5, a t-test reveals a significant difference in means on most of the relevant background characteristics of students. The Belgian sample is balanced across regions in terms of gender, age and parents' characteristics. However, important variables with respect to financial literacy, such as math and reading scores as well as the socioeconomic status in terms of wealth or the number of books at home, are significantly different in the two regional samples. PISA math scores, for example, are highly correlated with PISA financial literacy scores (in Flanders, $r(5673) = 0.793, p = 0.000$). Figure 19 shows that the distribution of math scores in Flanders and Wallonia overlaps; however the distribution is shifted to the right for Flanders compared to Wallonia. It is, therefore, likely that the financial literacy scores differ across the regions as well. A similar pattern is observed regarding the socioeconomic status of students in the two samples, as approximated by the PISA wealth indicator

[10] The Belgian federal state has three regions (Flanders, Wallonia, Brussels), but education policy is decided at the level of language communities (Flemish for schools in Flanders and Brussels, French for schools in Wallonia and Brussels, and German for schools in the German-speaking parts of Wallonia). To simplify, we group the French- and German-speaking communities and use the term Wallonia here as synonym for the French- and German-speaking communities and Flanders for the Flemish community. As of January 1st 2019, the German-speaking community account for 2% of the population of Wallonia.

[11] In the eight OECD countries and economies participating in the PISA financial literacy assessment 2015, this specific feature of differences in regional assessment is only the case for Belgium and Canada.

which summarises economic possessions of the family. Figure 19 shows that, similarly, the distribution of wealth overlaps, but is slightly shifted to the right in Flanders compared to Wallonia.

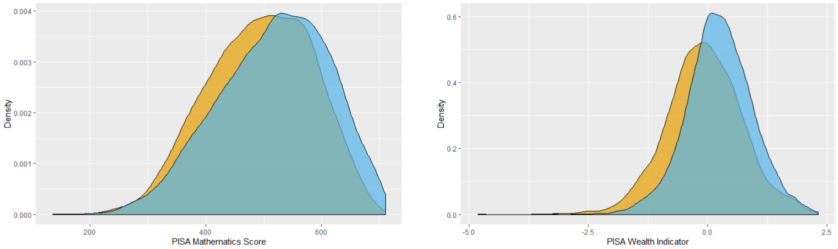


Figure 19: (Left) Distribution of PISA math Scores in Flanders (blue) and Wallonia (yellow).
(Right) Distribution of PISA Wealth Index in Flanders (blue) and Wallonia (yellow).

Having specific information about the financial literacy of students in each region is highly relevant for policy-makers in order to provide appropriate financial education. Predicting the missing financial literacy scores for Wallonia can therefore lead to informed policies tailored to the respective regional education system.

2.1 Outcome variable

The financial literacy scores in PISA 2015 are reported as a set of ten plausible values[12]. The financial literacy score is based on a test of financial knowledge with 43 items in the four content categories money and transactions, risk and reward, planning and managing finances, and the finan-

[12] For each participating country or economy, 33% of the students who completed the general PISA assessment were tested on financial literacy (OECD, 2017c). The plausible values were then constructed for all students participating in the general PISA based on item response theory and latent regression. In the case of Belgium, this was only done for Flanders. As a results, the plausible values are available for all participating students in the Flemish region, but missing for students from Wallonia. In the following, we use the plausible value $PV1FLIT$ for the analysis, since any bias caused by using a single plausible value instead of all ten simultaneously is arguably negligible (Gramațki, 2017).

cial landscape (OECD, 2017b).

In the case of Belgium, there are 9,651 observations on the general assessment from all Belgian regions, but the financial literacy scores are only available for the 5,675 students from Flanders, while the financial literacy scores of the 3,976 students from Wallonia are missing. Flanders thus represents our "study" population in which we observe the financial literacy scores Y^{obs}. Wallonia represents the "target" population for which we predict the missing financial literacy scores, Y^{obs}, based on the common set of predictors from the general assessment $\mathbf{X}$. Table 19 in Appendix 5 provides summary statistics of the outcome variable. On average, Flemish students score 541 points, which corresponds to the PISA proficiency level 3 out of 5 levels. Figure 27 in Appendix 5 shows the distribution of financial literacy scores in the Flemish data. 12% of Flemish students fail to reach the baseline level 2 of 400 points or more which the OECD defines as the level necessary to participate in society (OECD, 2017b). Figure 28 in Appendix 5 provides an overview of the required knowledge corresponding to the five PISA proficiency levels.

2.2 Predictors

In contrast to regression analyses, in which multicollinearity of regressors needs to be avoided, a large number of (potentially correlated) predictors can be used for forecasting in machine learning (Makridakis et al., 2008; Shmueli, 2010; Vaughan & Berry, 2005). We therefore select a broad set of predictors $\mathbf{X}$ from the general assessment of PISA 2015. Table 15 provides an overview of the variables used as predictors. The set of predictors includes students' background characteristics, proxies of the socioeconomic status of students, indicators of student achievement and attitudes, as well as school characteristics.

To account for the students' background, we include gender, grade, age, language and study track. Existing studies commonly find financial literacy scores to be highly correlated in particular with math and reading performance (e.g. Mancebón et al., 2019; Riitsalu & Põder, 2016). In the case of Flanders the correlation of financial literacy scores with

Table 15: Variables used from the PISA data

Student Characteristics	
ST001D01T	International Grade
ST004D01T	Gender
AGE	Age
ISCEDD	Study Track: ISCED Designation
ISCEDO	Study Track: ISCED Orientation
BELANGN	Speaks Belgian Language at Home
Socioeconomic Status	
HEDRES	Educational Resources at Home
WEALTH	Family Wealth Index (Economic Possessions)
ST013Q01TA	Number of Books at Home
IMMIG	Immigration Status
MISCED	Mother's Education (ISCED)
FISCED	Father's Education (ISCED)
BMMJ1	Mother's Job (ISEI)
BFMJ2	Father's Job (ISEI)
EMOSUPS	Parents Emotional Support
Achievement and Attitude	
PV1MATH	Plausible Value 1 in Mathematics
PV1READ	Plausible Value 1 in Reading
REPEAT	Grade Repetition
OUTHOURS	Out-of-School Study Time per Week
MMINS	Mathematics Learning Time at School
LMINS	Language Learning Time at School
ANXTEST	Personality: Test Anxiety
MOTIVAT	Achievement Motivation
School Characteristics	
SC001Q01TA	School Community (Location)
SC048Q01NA	Share of Students With a Different Heritage Language
SC048Q02NA	Share of Students With Special Needs
SC048Q03NA	Share of Socioeconomically Disadvantaged Students
SCHSIZE	School Size
CLSIZE	Class Size
RATCMP1	Number of Available Computers per Student
LEADPD	Teacher Professional Development
SCHAUT	School Autonomy
EDUSHORT	Shortage of Educational Material
STRATIO	Student-Teacher Ratio

math and reading scores amounts to 0.8, which is slightly higher than the OECD average of 0.75 (Universiteit Gent, 2017). As such, we include the PISA math score, as well as the reading score and additional variables related to student achievement, such as grade repetition, study time and instruction time. Given that personality traits and attitudes have been found to matter for financial literacy as well (Longobardi et al., 2018; Pesando, 2018), we also include students' test anxiety and motivation.

Another major factor associated with financial literacy scores is the family background, such as parental characteristics, language or immigration background (e.g. Gramaţki, 2017; Mancebón et al., 2019). To approximate the socioeconomic status of a student we use the indicators of educational and economic possessions of the family, the number of books in the home, the immigration background of the student, mother's and father's education and job, as well as a variable capturing perceived parental emotional support.

Finally, studies of PISA financial literacy scores commonly include variables for school level variables (e.g. Cordero & Pedraja, 2018; Pesando, 2018). We include a number of school characteristics from the questionnaire for school principals, such as the school's location, size and autonomy. As indicators of teaching quality, we use the student-teacher ratio, class size and teacher professional development. We also use school level indicators of socioeconomic status, such as the share of students with a different home language, special needs or a socioeconomically disadvantaged home, the number of available computers and shortage of educational material.

Figure 29 in Appendix 5 shows the missing observations in the variables used in the analysis. Apart from the financial literacy scores, which are, as described above, only available for Flemish students, no clear patterns of missingness appear across the predictors. We can therefore assume the observations to be Missing-Completely-at-Random (Little & Rubin, 2019) and we proceed with multiple imputations using the Fully Conditional Specification (FCS) developed by Buuren and Groothuis-Oudshoorn (2010) and implemented in the `R` package `MICE`.

3 Results

In this Section, we discuss the performance of BART in the prediction task at hand; we provide the results and some descriptive evidence regarding the observation with posterior predicted probability lower than the average; and we provide the results from the sensitivity analysis.

3.1 Results for BART predictions

Many recent studies have shown that BART performs in an excellent way in various predictive tasks across different scenarios (Bargagli-Stoffi, Riccaboni, et al., 2020; Hernández et al., 2018; Linero, 2018; Linero & Yang, 2018; Murray, 2017; Starling et al., 2018). In our particular scenario, we test the performance of BART through a 10-folds cross-validation, and compare it with the performance of the random forest (RF) algorithm Breiman (2001). A generic k-folds cross-validation technique consists of dividing the overall sample (in our specific case we draw this sample from the set of observations for which the outcome variable is present) into k different subsamples. $k-1$ subsamples are used to train the machine learning technique. Then its performance is tested on the subsample that was left out. We decide to test the performance of our algorithm with the one of the RF algorithm, one of the most widely used algorithms for prediction in social sciences (Athey, 2019; Bargagli-Stoffi, Niederreiter, et al., 2020). We observe that BART performs equally or better than the random forest algorithm in the predictive task at hand. Table 16 depicts the comparative results of the two algorithms with respect to their root-mean-squared-error (RMSE) and mean-absolute-error (MAE) of prediction and their R^2. BART demonstrates a very good performance (as well as a series of other qualities that were described briefly in Section 1), as it outperforms RF with respect to all the selected performance measures (i.e., smaller RMSE and MAE, higher R^2). Hence, we use this technique to predict the missing financial literacy scores for students in Wallonia.

Figure 20 depicts the posterior predicted values for financial literacy scores for students in Flanders (light blue) and Wallonia (orange), while

Table 16: Summary statistics of the performance of the ML techniques

	RMSE	*MAE*	R^2
BART	58.1799	45.7505	0.7306
Random Forest	58.8402	46.1451	0.7251

Table 17: Summary statistics of the predicted FLS

	Mean	SD	Minimum	Median	Maximum	N
Predicted FLS Flanders	541.4	96.8	157.3	559.9	750.0	5675
Predicted FLS Wallonia	516.6	95.7	188.3	530.3	731.7	3976

Table 17 reports summary statistics for the same predictions. From Table 17 we observe that the mean posterior predictive financial literacy scores for students in Flanders is higher than for those in Wallonia, 541.4 compared to 516.6. Considering the minimum and maximum values for Flanders and Wallonia we note that the scores of the students in Flanders seem to be more centered around the mean in Wallonia compared to Flanders. The same conclusion can be drawn from Figure 20 where the peak of the light blue area (Flanders) is to the right of the peak of the orange area (Wallonia). As OECD provide interpretable difference between the FLS, we can interpret the different results for Flanders and Wallonia. In particular, based on OECD computations a difference of 40 points equals an entire school year of learning. Hence, based on this, we can state that students in Flanders are roughly half a year ahead of students in Wallonia with respect to their standardized knowledge in financial literacy.

Robustness checks

The quality of the predictions of FLS in Wallonia is not directly testable as this outcome is unobserved in this region. However, the quality of these prediction can be thought to potentially be very similar to the one estimated through cross-validation if the underlying assumption that the 'technology' and the efficiency to transform teaching and environmental

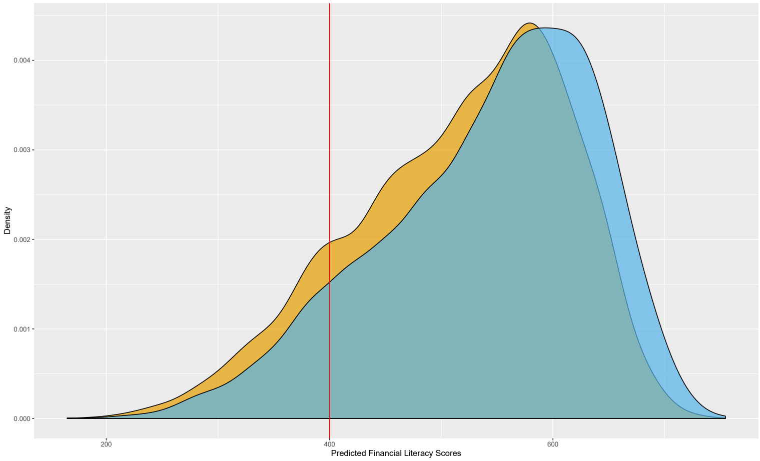


Figure 20: Predicted financial literacy scores for Flanders (light blue) and Wallonia (orange). The red line indicates the threshold of the baseline level of proficiency in financial literacy. OECD suggests that students above this threshold of 400 points have financial literacy levels that are sufficient to participate in society (OECD, 2017b).

inputs into learning output is comparable between Flanders and Wallonia. As we argued in Section 3, it is relatively safe to think about this assumption to hold.

However, to investigate this assumption thoroughly, we run a series of robustness checks using different outcome variables. Given the high correlation between the outcome and *math scores* and *reading scores*, and given the standardized way all these outcome are recorded, the most straightforward way to implement this analysis is to use these two variable as outcomes and test whether the BART model build using Flemish data performs well when tested on Walloon data.

To perform this test we use a reworked version of the 10-folds cross-validation used in the previous Section. This procedure is illustrated in Figure 21, and assigns random folds from the Flemish data to the training sample and one random fold from the Walloon data to the training sample at each split. The performance of the method is evaluated at each

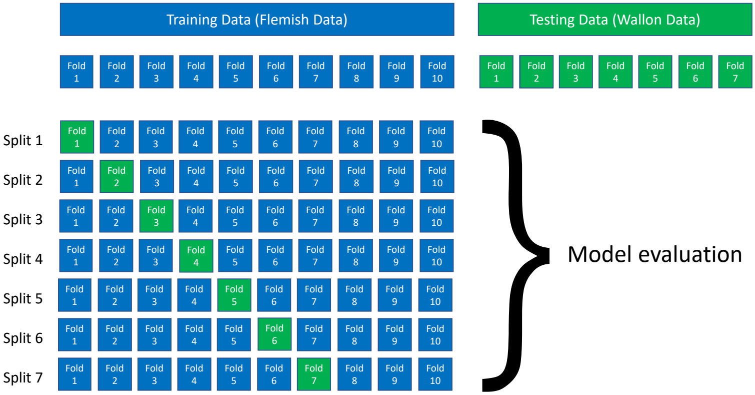


Figure 21: Reworked k-folds cross-validation.

split and then the results are aggregated. To make this example mimic as close as possible the evaluation of the algorithm performed in the previous Section, we chose to use the same number of training folds and testing fold. However, as the testing data are less then the training data (respectively 5,675 students from Flanders and 3,976 students from Wallonia), the training data were partitioned in just 7 folds and each fold was used at a different split to assess the performance of the algorithm.

The results are depicted in Table 18. The adjusted R^2 for BARTs built on the different outcomes are comparable to the ones in Table 16 and hint at the fact that there are no relevant differences between the performance of the model built using the Flemish data both for training and testing and the one built using Flemish data for training and Walloon data for testing. In turn, this robustness check seems to confirm the fact that the 'technology' and the efficiency in Flanders and Wallonia are comparable.

3.2 Results for vulnerable students detection

Policy-makers often want to target only those who are really in need for an intervention as they are often facing budget constraints that pre-

Table 18: Summary statistics of the performance of the BART on different outcomes

	RMSE	*MAE*	R^2
Math Score Outcome	52.5344	41.89435	0.7099
Reading Score Outcome	52.5421	441.8146	0.7064

vent them from enacting policies for all constituents. These groups can be identified, in cases where the values of a certain outcome of interest are observed, by simply looking at the outcome's distribution. However, it is often the case that such outcomes may be hidden or not observed for a part of the study population. In such scenarios, machine learning predictions may furnish a valuable tool to policy makers. In the case of our application, we face a prediction problem: imagine that policy makers want to provide additional learning material to more vulnerable students in a region where FLS are not observed. Machine learning can provide a useful tool to draw such predictions.

In order to detect the most vulnerable students (i.e., those with low predicted FLS), we use the procedure introduced in Section 4.1. Uncovering subpopulations of students that differ in terms of the distribution of a certain outcome is central to boost school and teacher's effectiveness and various strategies have been applied to this task (Masci et al., 2019). However, to the extent of our knowledge, this is the first time that a supervised machine learning technique is used for this aim. In a previous study on OECD data, De Beckker et al. (2019) identify four groups of adults with different financial literacy levels through an unsupervised machine learning algorithm (i.e., k-means). The advantage of the use of machine learning is clear as it allows to partition a large heterogeneous group of people into subgroups according to their financial literacy scores. However, while simple unsupervised machine learning algorithms such as k-means are useful to get a first impression of the different subgroups and their socio-economic characteristics, they do not provide immediate information on the robustness of the outcomes. Moreover, these simple algorithms are not capable of making predictions for out-of-sample regions. This is particularly relevant when dealing with standardized tests, such as PISA, which often only cover subpopulations of a country, such as specific regions.

First, we train BART on the sample of Flemish students for which we have data on their financial literacy scores. Next, we predict the financial literacy scores for both the sample of students in Flanders and Wallonia. After that, we compare the posterior predicted values for each student with the mean of the predicted posterior values. Finally, we detect the students for which these values are lying outside the credibility intervals for the mean values. Here, we define outliers as those observations with predicted financial literacy scores values smaller than two standard deviations from the mean following J. Miller (1991). In Appendix 6, we check the results we would obtain when defining the outliers as observations with predicted financial literacy scores two absolute deviations smaller than the median as suggested by Leys et al. (2013)[13].

Moreover, in order to get a better understanding of which variables best explain low predicted financial literacy scores we generate a dummy variable that assumes value 1 if the student has a low predicted financial literacy scores and 0 otherwise. Finally, we built a series of conditional inference trees (Hothorn et al., 2006) to have a descriptive sense of which are the drivers of low financial literacy. Figure 22 depicts the conditional tree for the overall sample, Figure 23 shows the conditional trees for Flanders and Wallonia respectively, Figure 24 depicts the trees for students in grades 7 to 9 and 10 to 12, and Figure 25 shows the results for the trees for students in general and vocational education.

As shown in Figure 22, reading (PV1READ) and math (PV1MATH) scores, the language spoken at home (BELANGN), study track (ISCEDD) and grade (ST001D01T) are important variables distinguishing groups of students in our entire sample. For students from the lower grades (grade 7 to 9) with lower scores on reading (≤ 389.567) and math (≤ 406.49) the predicted probability of a low financial literacy scores is 95 percent. A

[13] Such results are not significantly different from the ones that are depicted in this Section. Thus, we argue that our results are robust to different definitions of outlying predictions.

higher PISA math score (> 406.49), but not being a native speaker results in a predicted probability of 84 percent of having a low financial literacy scores. On the opposite side, students with a reading score above baseline (> 389.567), following general programs designed to give access to the next level, have - depending on their grade - a predicted probability of respectively only 19 or 1 percent of having a low financial literacy scores.

In Figure 23, we split by region. In the tree in the left panel of the Figure, we note that in Flanders the largest predictive proportion of students with low financial literacy scores are again situated among those with lower scores for reading (≤ 386.732) and math (≤ 405.291). In the group with a low reading score (≤ 386.732) and a higher math score (> 405.291) immigration status seems to matter (IMMIG). Among those who are first generation-migrants the predictive probability of having a low financial literacy scores is 82 percent. In Wallonia (tree in the right panel), the proportion of students with low financial literacy scores is again the largest (94 percent) among those with low reading (≤ 385.283) and math scores (≤ 406.49). In Wallonia, mother's education (MISCED) seems to play a distinctive role. Students with better reading and math scores and with a highly educated mother only have a predictive probability of 1 percent of low financial literacy scores.

In Figure 24, we split by grade. We group grades 7-9 (left panel), and 10-12 (right panel). Indeed in Flanders, after grade 9 certain conditions change (teacher qualification requirements, etc.). Hence, it is interesting to investigated the drivers of heterogeneous effects by grade. In grades 7-9, the largest predictive proportion (95 percent) of students with low financial literacy scores are again among those with a lower reading (≤ 398.29) and math score (≤ 389.567). The second largest proportion of students with low financial literacy scores (84 percent) is found among non-native speakers with a high math score (406.49) but with a low reading score (≤ 398.29). The same pattern seems to exists in grades 10 to 12. However, in grades 10 to 12, the education of a student's father (FISCED) seems to have a significant influence on financial literacy scores. For students with fathers with a higher education background, the predictive probability of low financial literacy scores lies between 1 and 20 percent, depending on whether the student has a math score above or below 399.714.

In Figure 25, we split by track. In the case of general education students (left panel) from grade 7, 8 and 12 with a low reading (≤ 397.528) and math score ($\leq$ 409.921) have, with a predictive probability of 92 percent, the highest chance of having low financial literacy scores. The second largest predictive proportion (78 percent) is among students in grade 9, 10 and 11 in schools with more than 67 percent of special needs students (SC0848Q02NA) who have a math score smaller or equal to 460.087. Considering only the students in vocational education (right panel), we observe that the main determinants of students' low financial literacy are the study track, math and reading scores and whether or not the student speaks Dutch at home.

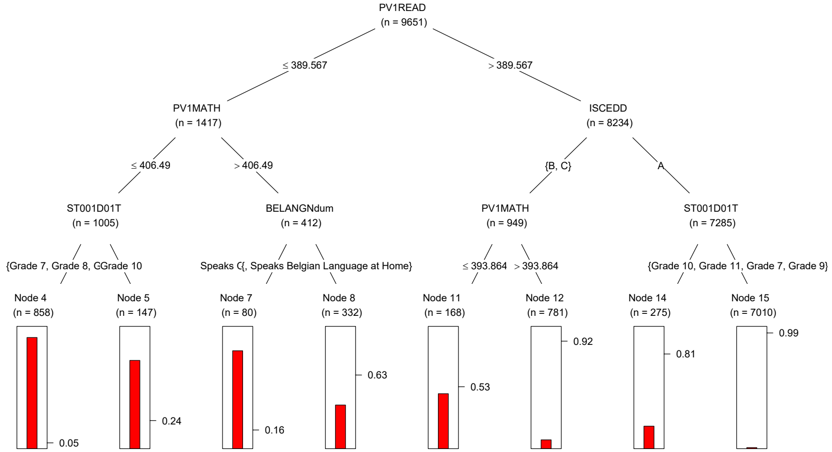


Figure 22: Conditional tree for the entire sample. Within each leaf is depicted in red the histogram of the percentage of units that have a low financial literacy score, and next to it the percentage of units with not-low financial literacy score within the same leaf.

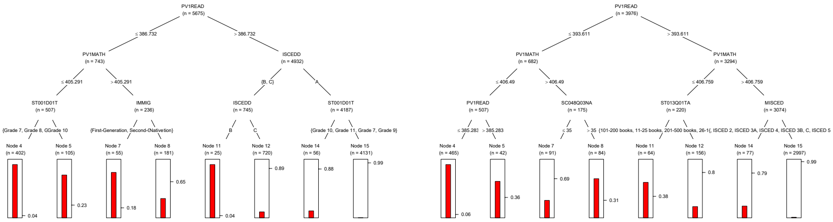


Figure 23: (Left) conditional tree for Flanders. (Right) The corresponding tree for Wallonia.

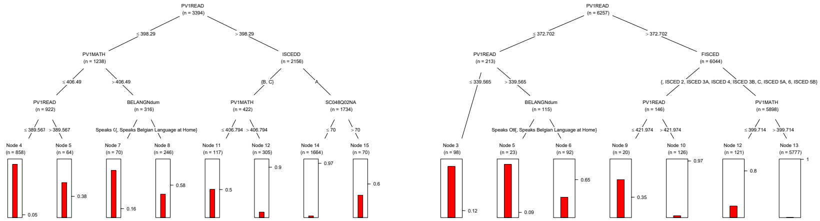


Figure 24: (Left) conditional tree for students in grades 7 to 9. (Right) The corresponding tree for students in grades 10 to 12.

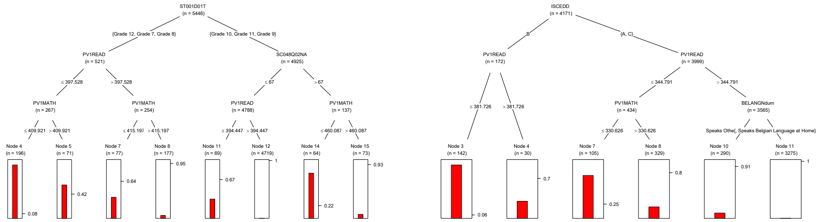


Figure 25: (Left) conditional tree for general education. (Right) The corresponding tree for vocational education.

3.3 Results for sensitivity of predictions analysis

In this Section, we report the results for the sensitivity of predictions analysis. In Section 1.2 we introduced the methodology which we employ here to assess how an unobserved predictor, uncorrelated to the observed ones, could impact the predicted financial literacy scores and the performance of the model. In particular, the `sensitivity` function that we developed in the statistical software `R` provides estimates of how many unit level predictions are statistically different between the original model and the model with the synthetic predictor and how much the synthetic predictor would improve the overall performance of the model.

Starting from the unit level predictions that we introduced in Section 1.2, we can get a sense of how much the synthetic predictor would affect the model prediction in a varying range of correlations between the synthetic predictor and the outcome (in the case of our application this range is $[0.1, 0.5]$)[14]. Here, we test how many of the predicted values of the augmented model would be statistically different from the predicted values of the original model. We find no evidence of statistically significantly different predictions up to a correlation of 0.5, where just 0.01% of the unit level predictions of the augmented model are significantly different.

Moreover, figure 26 depicts the results for both the RMSE and R^2 of the original and augmented model with their confidence intervals as the correlation between the synthetic predictor R and the outcome Y increases. In both cases, even when the synthetic predictor has a correlation of 0.5 with the outcome, there is no significant difference between the RMSE of the original and the augmented model. The R^2 is significantly improved just when the correlation reaches 0.5. This evidence could be interpreted as the fact that, in order to get a significant increase in the R^2, we would need an explanatory variable (completely unrelated to the observed ones) with at least a 0.5 correlation to the outcome. To put things into perspective, a predictor with such a correlation would rank between the first three best predictors in the model. We argue that, in the context

[14] This range depends on the data at hand as we need to guarantee that $\mathbf{R}$ is a symmetric positive definite matrix.

of our application, such an unobserved variable would be very hard, if not impossible, to find.

As argued before, the interpretation of such a sensitivity analysis is that, even if we could introduce a new, unobserved predictor, with a high explanatory power in the model, the unit level predictions as well as the overall performance of the model would remain fairly stable. This set of evidence, together with the non-significant variation in the unit level predictions, hints at the fact that our model is already capturing most of the signal in the data and its predictions are stable. Hence, this analysis offers to policy-makers an additional evidence in favour of using the predictions from the model that we propose to enact targeted policies.

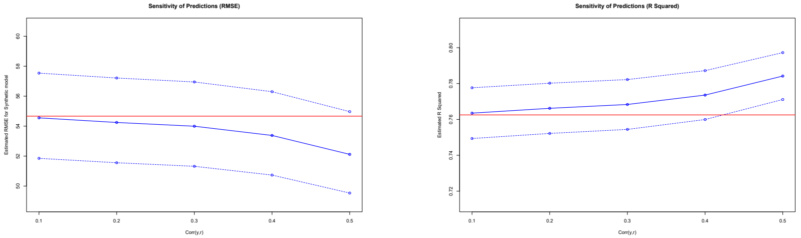


Figure 26: (Left) Estimated RMSE for the augmented model with confidence intervals (blue) and for the original model (red). (Right) Estimated R^2 for the augmented model with confidence intervals (blue) and for the original model (red). The 99% confidence intervals were constructed following Cohen et al. (2014).

Moreover, we make this analysis robust by introducing correlations between the synthetic predictor and the two best predictors The size of the correlation is the same as the correlation introduced between the synthetic predictor and the outcome. In the case of our application, the model is robust because the results from the sensitivity analyses are invariant to the introduction of this additional correlation patterns.

[14] The best predictors are chosen according to the built-in variables' importance measure of BART. For a review of various tree-based variables' importance measures and their pitfalls we refer the reader to James et al. (2013) and the recent contribution by Gottard et al. (2020).

4 Discussion

In this Chapter, we introduce a novel sensitivity analysis to assess the robustness of our preferred predictive model in terms of the stability of its predictions and its performance. The model that we propose is general enough to be applied to both Bayesian techniques (such as the Bayesian Additive Regression Trees algorithm) or frequentist techniques (such as the Random Forest algorithm). In particular, we propose a novel way to partially answer policy-relevant questions such as *"is the machine learning model getting enough signal from the data?"* and *"how much would an unobserved predictor impact the model's prediction and its performance?"*. To do so, we develop a novel methodology to construct a synthetic predictor that is correlated with the outcome but is uncorrelated with the observed predictors. By generating the synthetic predictor in this a way, we are able to interpret the results from the sensitivity analysis in a way that hints at how much an exogenous predictor with a high explanatory value would add to the model. If the addition of the synthetic predictor is not resulting in significant differences, then we can safely assume that our predictive model is capturing enough signal from the training data. Moreover, through our novel methodology we are able to use a wide range of correlations between the outcome and the synthetic predictor. This enables the researchers to check whether or not there is a *break-point* (given by the explanatory power of the synthetic predictor) above which the performance of the model would be significantly better.

In addition, we use a novel machine learning approach to predict financial literacy scores for students in a region of Belgium where financial literacy scores are unobserved. We identify the characteristics of students with predicted financial literacy scores lower than the mean. Finally, we verify the sensitivity of our observations for unobserved predictors.

The first stage of our analysis provides evidence that the Bayesian Additive Regression Tree (BART) approach outperforms the traditional random forest (RF) approach on a number of selected performance measures (i.e. smaller RMSE and MAE, higher R^2) with respect to FLS predictions. We find that the predicted values for financial literacy scores for students in Flanders are on average somewhat larger than those for students in Wallonia. Nevertheless, the distribution of financial literacy scores in Flanders is more extreme (i.e. lower minimum, larger maximum) compared to Wallonia.

Next, we estimate the posterior predicted observations available for each student which also allows to identify outliers (i.e. students for which the the financial literacy scores are situated outside the 95% credibility intervals for the mean values). Our results show that the predictive probability of having a low financial literacy scores is the largest for students with lower scores on reading and math in the PISA test. This corroborates with findings of Mancebón et al. (2019) who state that the development of financial abilities of students is mediated by their mathematical skills. Another interesting observation is that the family background also has a key influence. Students with the largest predicted low financial literacy scores are often students from families where the school language is not spoken at home. Particularly in Flanders, being a first-generation immigrant or not is key in having low financial literacy scores. The educational background of parents is another important predictor. These observations are in line with existing literature suggesting that more vulnerable groups in terms of low financial literacy scores are often situated among those with a lower socioeconomic background (De Beckker et al., 2019; Gramaţki, 2017; Riitsalu & Põder, 2016).

Finally, we confirm that our novel approach is robust in terms of potential unobserved variables. The inclusion of a synthetic predictor highly correlated with the outcome does result in a significant difference in the performance measures of the original model and the synthetic model. From a policy perspective, our novel approach is highly interesting as it not only allows to predict outcomes for certain countries or regions within countries with missing data, but also introduces the possibility to detect observations that fall outside the mean outcomes. In this respect, policy-makers can detect which factors drive low results. We applied our model to PISA financial literacy data, however, the same approach can also be applied to other large administrative datasets.

Supplementary material for Chapter 3

5 Data

Summary statistics of Plausible Value 1 in Financial Literacy from PISA 2015 for the Flemish region in Belgium. The OECD constructs ten plausible values for financial literacy using item response theory and latent regression (OECD, 2017c). The analysis in this Chapter is based on Plausible Value 1. SD stands for the standard deviation of the variable.

Table 19: Summary statistics of the outcome variable

	Mean	SD	Minimum	Median	Maximum	N
Financial Literacy	541.43	112.16	51.81	555.14	901.64	5675

6 Changing outliers definition

Below we depict the results for the conditional tree analysis, when we define outliers as observation with predicted values below two absolute deviation from the median. This is a more restrictive definition of outliers.

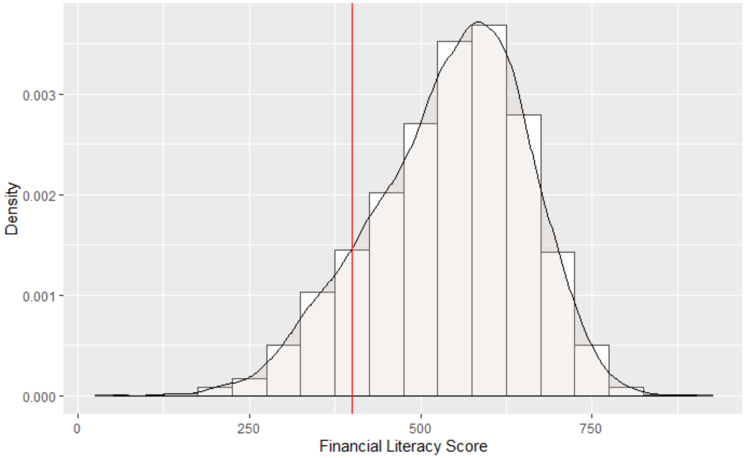
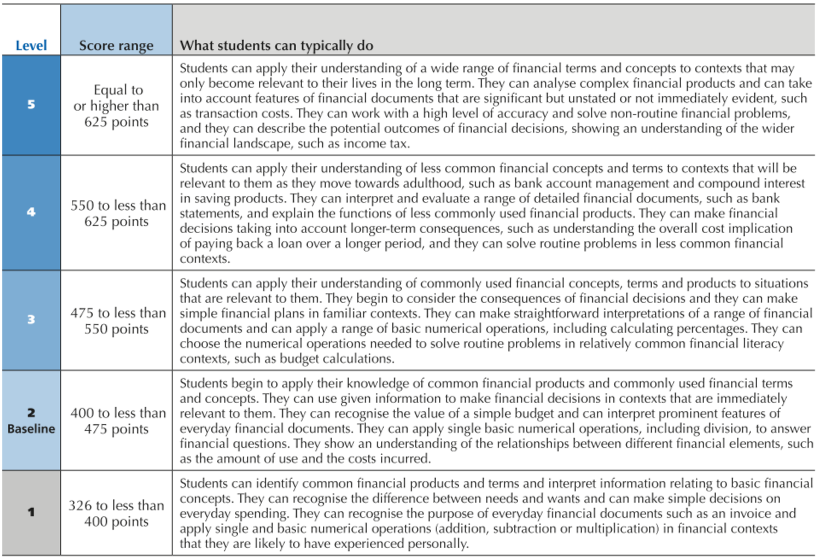


Figure 27: Histogram of the Outcome Variable (PISA Financial Literacy Score). The red line indicates the threshold of the baseline level of proficiency in financial literacy. The OECD suggests that students above this threshold of 400 points have financial literacy levels that are sufficient to participate in society (OECD, 2017b).



Level	Score range	What students can typically do
5	Equal to or higher than 625 points	Students can apply their understanding of a wide range of financial terms and concepts to contexts that may only become relevant to their lives in the long term. They can analyse complex financial products and can take into account features of financial documents that are significant but unstated or not immediately evident, such as transaction costs. They can work with a high level of accuracy and solve non-routine financial problems, and they can describe the potential outcomes of financial decisions, showing an understanding of the wider financial landscape, such as income tax.
4	550 to less than 625 points	Students can apply their understanding of less common financial concepts and terms to contexts that will be relevant to them as they move towards adulthood, such as bank account management and compound interest in saving products. They can interpret and evaluate a range of detailed financial documents, such as bank statements, and explain the functions of less commonly used financial products. They can make financial decisions taking into account longer-term consequences, such as understanding the overall cost implication of paying back a loan over a longer period, and they can solve routine problems in less common financial contexts.
3	475 to less than 550 points	Students can apply their understanding of commonly used financial concepts, terms and products to situations that are relevant to them. They begin to consider the consequences of financial decisions and they can make simple financial plans in familiar contexts. They can make straightforward interpretations of a range of financial documents and can apply a range of basic numerical operations, including calculating percentages. They can choose the numerical operations needed to solve routine problems in relatively common financial literacy contexts, such as budget calculations.
2 Baseline	400 to less than 475 points	Students begin to apply their knowledge of common financial products and commonly used financial terms and concepts. They can use given information to make financial decisions in contexts that are immediately relevant to them. They can recognise the value of a simple budget and can interpret prominent features of everyday financial documents. They can apply single basic numerical operations, including division, to answer financial questions. They show an understanding of the relationships between different financial elements, such as the amount of use and the costs incurred.
1	326 to less than 400 points	Students can identify common financial products and terms and interpret information relating to basic financial concepts. They can recognise the difference between needs and wants and can make simple decisions on everyday spending. They can recognise the purpose of everyday financial documents such as an invoice and apply single and basic numerical operations (addition, subtraction or multiplication) in financial contexts that they are likely to have experienced personally.

Figure 28: PISA Proficiency Levels for Financial Literacy (OECD, 2017b)

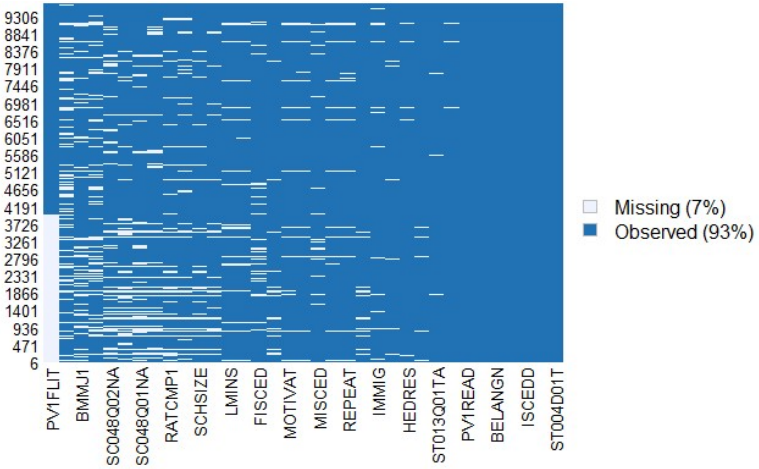


Figure 29: Missingness map

Table 20: Summary statistics of the predictors

	Flanders			Wallonia			Difference in Means
	Mean	SD	N	Mean	SD	N	p-value
Student Characteristics							
International Grade	9.72	0.52	5675	9.44	0.75	3857	0.000
Gender	1.51	0.50	5675	1.51	0.50	3976	0.976
Age	15.85	0.29	5675	15.84	0.29	3976	0.725
Study Track: ISCED Designation	1.43	0.81	5675	1.24	0.64	3976	0.000
Study Track: ISCED Orientation	2.09	1.00	5675	1.55	0.89	3976	0.000
Speaks Belgian Language at Home	0.91	0.29	5595	0.87	0.34	3929	0.000
Socioeconomic Status							
Educational Resources at Home	0.27	0.91	5585	-0.15	0.90	3930	0.000
Family Wealth Index (Economic Possessions)	0.27	0.74	5606	-0.04	0.85	3936	0.000
Number of Books at Home	3.02	1.51	5557	3.32	1.53	3905	0.000
Immigration Status	1.20	0.54	5512	1.32	0.66	3851	0.000
Mother's Education (ISCED)	4.62	1.39	5376	4.63	1.53	3774	0.882
Father's Education (ISCED)	4.53	1.44	5221	4.52	1.58	3654	0.901
Mother's Job (ISEI)	47.05	22.75	4763	47.30	22.10	3159	0.635
Father's Job (ISEI)	46.21	21.38	4713	46.21	22.16	3318	0.996
Parents Emotional Support	-0.01	0.96	5458	0.01	0.96	3801	0.302
Achievement and Attitude							
Plausible Value 1 in Mathematics	521.61	97.89	5675	494.78	90.71	3976	0.000
Plausible Value 1 in Reading	510.95	100.67	5675	490.46	95.81	3976	0.000
Grade Repetition	0.24	0.43	5457	0.42	0.49	3804	0.000
Out-of-School Study Time per Week	14.24	10.21	4245	16.27	11.61	3216	0.000
Mathematics Learning Time at School	191.28	86.43	5273	219.74	81.86	3668	0.000
Language Learning Time at School	187.24	84.63	5275	227.10	83.49	3665	0.000
Personality: Test Anxiety	-0.30	0.97	5416	-0.01	1.01	3775	0.000
Achievement Motivation	-0.64	0.83	5417	-0.28	0.87	3767	0.000
School Characteristics							
School Community (Location)	2.87	0.81	5556	3.24	1.11	3681	0.000
Share of Students With a Different Heritage Language	16.41	23.39	5406	26.19	30.29	2812	0.000
Share of Students With Special Needs	19.58	20.14	5154	18.97	21.94	3005	0.208
Share of Socioeconomically Disadvantaged Students	20.02	22.11	5248	33.06	30.38	3120	0.000
School Size	695.86	331.42	5405	771.27	327.30	3413	0.000
Class Size	18.89	5.14	5592	21.01	3.77	3578	0.000
Number of Available Computers per Student	1.25	0.89	5270	0.47	0.33	3251	0.000
Teacher Professional Development	0.10	0.90	5150	0.11	1.07	3433	0.685
School Autonomy	0.77	0.18	5675	0.59	0.21	3626	0.000
Shortage of Educational Material	0.02	0.87	5488	0.21	0.87	3429	0.000
Student-Teacher Ratio	9.09	3.21	5325	9.14	2.66	2847	0.454

Note: Summary statistics of all predictors used in the analysis from the PISA 2015 data for Flanders and Wallonia. SD stands for the standard deviation of the variable. The last column shows the p-value of a two-sample t-test for the equality of means across the regions of Flanders and Wallonia. Language has been grouped for Belgium (1=Dutch, 2=French, 3=German, 4=Other).

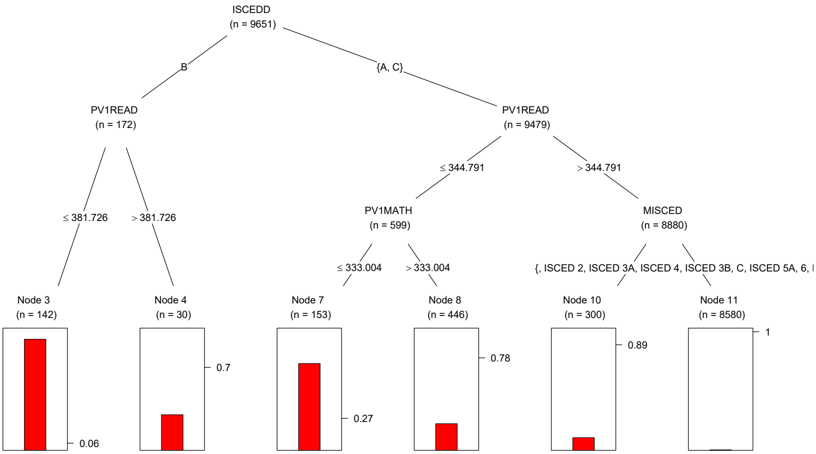


Figure 30: Conditional tree for the entire sample. Within each leaf are depicted in red the histogram of the percentage of units that have a low financial literacy score, and next to it the percentage of units with not-low financial literacy score within the same leaf.

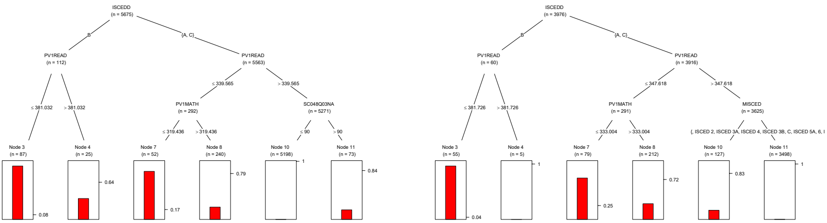


Figure 31: (Left) conditional tree for Flanders. (Right) The corresponding tree for Wallonia.

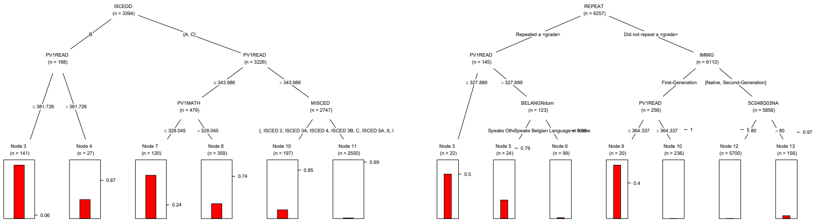


Figure 32: (Left) conditional tree for students in grades 7 to 9. (Right) The corresponding tree for students in grades 10 to 12.

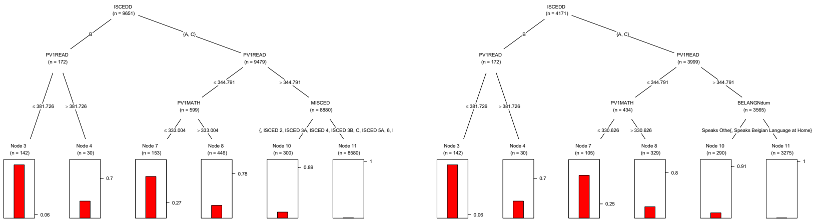


Figure 33: (Left) conditional tree for general education. (Right) The corresponding tree for vocational education.

Chapter 4

Machine learning for zombie hunting

In this Chapter, we use machine learning techniques to identify *zombie firms* after training an algorithm on a big dataset of firms' failures and financial accounts. We implement a Bayesian Additive Regression Tree with Missing Incorporated in Attributes (BART-MIA) robust to non-random patterns of missing values to predict firms' failures. A LASSO (least absolute shrinkage and selection operator) allows us selecting a list of best predictors from firm level observed information. Eventually, we propose an empirical definition of *zombie firms* as firms that persist in a status of high risk of failure.

The problem of spotting non-viable firms is relevant for both academics and practitioners, whether the reason is to assess credit risk for the individual firm or, more in general, to detect a share of an economy that is in trouble. On one hand, we have a large body of literature that proposes to assess credit risk from the perspective of a financial institution. On the other hand, we have an emerging literature that assesses the viability of firms from a wider perspective, possibly discussing policies that reduce the misallocation of financial and productive resources, as for example in the case of *zombie firms*. In either cases, we argue that

This Chapter is based on Bargagli-Stoffi, Riccaboni and Rungi (2020).

machine learning techniques improve upon existing proxy methods and standard econometric techniques, reducing inescapable prediction errors while paving the way for more target-specific and evidence-based policies.

From the perspective of a financial company, a limited information is usually available on the viability of firms' investment projects from its financial accounts. The seminal references in this field are the Z-Score (Altman, 1968, 2000) and the distance-to-default (Merton, 1974) proxy models. In both cases, firms' financial accounts have been traditionally used to gauge a company's likelihood of bankruptcy. In the case of the Z-Score, five financial ratios are calculated from firm-level data, typically including profitability, leverage, liquidity, solvency and volumes of activity. After the latter are plugged in an equation with some weights to proxy their relative importance, one obtains a threshold that, if crossed, indicates that there is a high probability of bankruptcy in the next future. Different from the Z-Score, the distance-to-default model by Merton (1974) focuses specifically on the ability of a company to meet its financial obligations. The original intuition is that a company's equity can be modelled as a call option on its assets. Hence, the distance-to-default also requires the knowledge of some firm-level accounts (company's assets, debts, market value) combined with information retrieved on financial markets (risk-free interest rate, standard deviation of stock returns) to plug in an equation that returns the value of a theoretically fair call option[1].

Against this background, our machine learning technique clearly improves upon existing models of credit rating by exploiting a wider battery of firm-level economic and financial indicators that potentially contain different pieces of information regarding both the core economic activity of the firm and its ability to meet financial obligations. In fact, by construction, machine learning techniques work better if one includes as much valuable information as possible, and they are dynamic as they

[1] After the insights of the distance-to-default model, Black and Scholes (1973) developed their widely known model based on the observation that hedging an option one could remove a systemic risk component.

continuously update themselves every time there is new out-of-sample information. The more we update the prediction algorithm the more precise the prediction we obtain. Moreover, statistical learning allows controlling for the level of significance of the predictions, and it provides us with the possibility to measure how prediction errors reduce thanks to independent tests that minimize deviations between realized and predicted outcomes (Athey, 2019). In fact, we argue that any closed set of indicators, any *ex-ante* weights attributed to ratios, and any specific functional forms that we find in previous proxy methods are not flexible enough to consider the ever-changing peculiarities of the economic environment in which each company operates. One size does not fit all.

On the contrary, we can explicitly track non-parametrically whether and how some firms' accounts become more or less relevant than others to predict failure over time. For example, the same firms' accounts can show more distress in times of a general crisis than in happier times. In addition, firms in one country could be more resilient than others, although under distress, as a result of some institutional characteristics.

As we acknowledge that we will never be able to expunge uncertainty and informative asymmetries from our horizon, we can instead reduce our disadvantage after using all the information at our disposal with iterative statistical learning methods. After we apply a BART-MIA on a sample of 304,906 Italian firms in the period 2008-2017, we show how it outperforms previous proxy models (i.e., Z-score and distance-to-default), standard econometric methods (i.e., a logistic regression), and also other machine learning techniques (i.e., Classification and Regression Tree, Random Forest, Super Learner) in predicting firms' failures. From our point of view, Italy is an interesting case as a country that has been shown to host an important share of non-viable and less productive firms that hamper the growth potential of the economy (Andrews et al., 2017; Calligaris et al., 2016; McGowan et al., 2018). Eventually, we end up with a probabilistic measure for the financial healthiness of a firm, which can be used to assess a firm's credit risk. The higher the value of the probabilistic measure the less viable the firm is on the market.

Among indicators that better forecast failure, we find the firm-level ratio between interest payments and cash flow, proposed as a financial constraint indicator by Nickell and Nicolitsas (1999), but also the interest coverage ratio, a negative value added, a size-age indicator proposed by Hadlock and Pierce (2010), a profitability indicator (Schivardi et al., 2017), a benchmark difference with firm-level interest payments (Caballero et al., 2008; McGowan et al., 2018), the liquidity and solvency ratios. Yet, we underline how no single indicator can be separated from the entire battery to predict failure, because each taken alone ends up with a higher rate of *false positives*, when the test wrongly indicates that a firm is at a high risk of failure, or a higher rate of *false negatives,* when the test wrongly indicates that a firm is viable.

From our point of view, the empirical problem of assessing if a firm is a *zombie* is strictly related to the problem of assessing its credit risk and its resilience to bankruptcy. In either case, a company has private information on the actual viability of its investment projects that is not disclosed to the financial institution or, more in general, to an analyst that may want to understand how much of a portion of an economy is in trouble. In either case, such a private information may be only in part observed from financial accounts. In the end, we argue that a *zombie firm* is a non-viable firm that is able to escape failure thanks to external financial resources despite its inability to repay debts. In other words, *zombies* are firms that have the highest risk of failure but they do not fail thanks to external financing. Interestingly, most of the indicators that we have found to be good predictors of failure have also been used to spot *zombies*, each of them in a different framework. Recent OECD studies (Andrews et al., 2017) estimate that so-called *zombies* represent a non-negligible and increasing share of many economies in a range between 2% and 10% of incumbent firms in year 2013, and they absorb up to 15%, 19% and 28% of the capital stock in countries like Spain, Italy and Greece, respectively. Eventually, in some economic environments more than in others, market selection processes seem not be working properly for a variety of reasons, and spotting non-viable firms could be especially useful for avoiding a waste of economic and financial resources. In this contribution, we set a working threshold of high risk of failure over the 9th decile and assume that a *zombie* is a firm that persists in such a high risk status. In the Italian case, we find that the share of *zombies* falls in a range between 2.5% and 4.5%. Interestingly, we also find that the share of *zombies* is countercyclical, because it increases during the financial crisis after 2008, but it decreases again after the economic recovery started.

We finally argue that spotting non-viable firms is important not only for financial institutions, but it is also of more general interest to the policy maker, as for example when she has to design optimal bankruptcy laws[2]. An empirical method that tracks a firm's risk of failure allows all stakeholders, not only creditors, to understand whether there is a chance for restructuring and, if not, to prevent that additional economic resources are wasted by incumbent albeit non-viable firms.

The rest of the Chapter is organized as follows. In Section 1, we introduce our empirical strategy for the identification of a firm-level risk of failures, and how it relates to the phenomenon of *zombies*. In Section 2, we introduce a novel indicator to spot *zombie firms* and we develop a methodology to assess which are the best predictors of *zombies*. In Section 3, we introduce our data and we provide some preliminary evidence. In Section 4, we discuss our predictions and related sensitivity checks. Section 5 concludes the Chapter with a discussion of potential applications and future work.

1 Empirical strategy

Spotting a non-viable firm is difficult for obvious reasons. If financial accounts are bad, one could always argue that it is just a matter of time because the firm will become more competitive in the next future, given the right conditions. If financial accounts are good, one could argue that the worst has yet to come and bad management choices will show up only later, given the right conditions. Trivially, only firms that are already bankrupt were certainly non-viable at some point, but an external

[2] See for example the European Directive 2012/30/UE and the recent Italian Law on business failures on October 19th 2017, n. 155 that laid the legal foundations for early signaling of firm's crises to enhance targeted interventions.

observer will never know *ex ante*, because a manager of a firm in trouble does not share her private information. This is the typical twin problem of a financial institution that faces uncertainty in presence of informative asymmetries. On one hand, the company under financial assessment has an obvious information advantage on its investment plans. On the other hand, both the financial institution and the company itself have a limited power in predicting future economic shocks that will have either positive or negative impacts on financial accounts. In this context, we propose a machine learning framework that uses past information about already failed firms to assess what the probability is that another firm in a similar condition will go bankrupt[3]. The wider the set and variety of past experiences on which we can rely, the more precise the prediction on the healthiness, or lack thereof, of a firm (Kleinberg et al., 2015). Ultimately, ours is a perspective on the (lack of) resilience of a firm potentially using all the observables that could possibly hold a piece of information on the viability of a firm. We end up with a firm-level probabilistic measure bounded between 0 and 1 that tells us what is a chance that a firm exits from the market in the next period since other firms in similar conditions did. Importantly, we are not concerned about the identification of the drivers of failures because ours is a pure prediction problem. Nonetheless, we do provide a methodology to assess a list of best predictors of which firm is *zombie* (accordingly to our working definition), possibly changing over time, which is not the same as finding the causes of failure. In fact, we show and discuss how predictors of failures can change over time and in different economic environments.

Finally, we discuss how the identification of *zombie firms* (Bank of England, 2013; Bank of Korea, 2013; Caballero et al., 2008; McGowan et al., 2018; Schivardi et al., 2017) can be usefully framed following the same prediction problem, hence providing an empirical definition of zombies

[3] Previous exercises to predict firms' failures from financial accounts have been tried in data science literature in absence of a clear economic framework. In this regard, Bargagli-Stoffi, Niederreiter, et al. (2020), provide an extensive overview of the recent data science literature related to the usage of machine learning for firm dynamics. The prediction power of our framework outperforms those exercises, as they are confined in a range 85% to 93% of Area Under the Curve (AUC)

as persisting in a status of high risk of failure. In the following analyses, we set a working threshold of high risk above the 9th decile of the distribution of fitted values, therefore we consider persistence when the firm outlives the market for at least three years without any sign of financial recovery.

1.1 Predictions of failures

Let us consider a generic prediction $\hat{f}$ in the form:

$$\hat{f}(\mathbf{X}_{it-1}) = \hat{P}(Y_{it} = 1 \mid \mathbf{X}_{it-1} = x_i), \tag{4.1}$$

where Y_{it} is the binary variable that assumes value 1 if the ith firm fails at time t and zero otherwise, while $\mathbf{X}_{it-1}$ is a matrix of firm-level predictors taken at time $t-1$. The functional form that links predictors to outcomes is determined by the generic supervised machine learning technique that will predict well out-of-sample. Briefly, the generic algorithm will pick the best in-sample loss-minimizing function in the form introduced in the Introduction section of the present Dissertation:

$$argmin \sum_{i=1}^{N} L(f(x_i), y_i) \quad over \quad f(\cdot) \in F \qquad s.\,t. \qquad R\big(f(\cdot)\big) \leq c, \tag{4.2}$$

where F is a function class from where to pick $f(\cdot)$, and $R\big(f(\cdot)\big)$ is the generic regularizer that summarizes the complexity of $f(\cdot)$. In our case, the function $f(\cdot)$ is an element of the family of classification trees, or a combination thereof. The set of regularizers, R's, will change following the standards adopted by each algorithm. Eventually, any algorithm we use shall take a loss function $L(\hat{y}, y)$ as an input and look for the function that minimizes prediction losses[4].

[4] In general, any tree $|\mathbb{T}|$ is built on *if-then* statements that split the dataset according to the observed value of the predictors, allowing for non-linear relationships between the predictors and the outcome. Hence, the generic algorithm for the construction of a tree, $\mathbb{T}$, is based on a top-down approach that continues to split the main sample into non-overlapping sub-samples (i.e. the nodes and the leaves). Then, the tree is pruned iteratively with a generic regularizer R to improve its predictive ability and avoid data overfitting, when trees become too deep over many layers. For a general reference, see Breiman et al. (1984).

Eventually, given new out-of-sample information, we can derive a prediction for the failure of any firm based on the set of its up-to-date financial accounts, both in the case of incumbent firms that already operated in the previous periods and in the case of firms entering into the market for the first time. From another perspective, we interpret this probability range as a degree of risk by an investor that does not have any other information than what is included in the actual financial accounts.

In following analyses, we consider as a baseline the BART - Bayesian Additive Regression Tree (Chipman et al., 2010). More specifically, we adopt a variant (BART-MIA) that is robust to non-random patterns of missing values in financial accounts to predict firms' failures (Kapelner & Bleich, 2015).

In general as we saw in Chapter 1, the BART (Chipman et al., 2010) is an ensemble method based on the aggregation of different independent trees and three regularizing priors designed to prevent overfitting. BART-MIA extends the original BART algorithm by incorporating additional information coming from patterns of missing values (Kapelner & Bleich, 2015). This is done by introducing, in each binary-tree component of the BART algorithm, the possibility to split on a missingness feature. As shown by Twala et al. (2008), such a splitting-rule allows trees to better capture the direct influence of missing values as a further predictor of the response variable. Indeed, in our framework, we can assume that patterns of missing values are not at random, or at least not completely at random, when firms are financially distressed. In other words, at least a few firms can try and hide private information by not reporting financial accounts, especially when they are in trouble. This is particularly true when some accounts are not legally binding. In this case, naïvely trimming the missing observations would introduce a sample selection bias in the data, i.e. we could more likely exclude from our analyses some firms that are financially distressed. Neither standard imputation techniques (e.g., conditional median imputation, Bayesian imputation) would solve such a sample selection problem, as far as they are based on the use of in-sample information. Usefully, the BART-MIA creates a new label when the values of a certain firm-level financial account is missing and it tests the new label as an additional predictor for a firm's failure.

1.2 Robustness checks

We compare baseline predictions from BART-MIA with other commonly used machine learning techniques: the Conditional Inference Tree (Hothorn et al., 2006), the Random Forest (Breiman, 2001), and the Super Learner (Van der Laan et al., 2007). Moreover, machine learning predictions are compared with a standard logistic regression model, as a benchmark for classical econometric techniques. Finally, we contrast our predictions with Z-scores and proxy models obtained from Altman (1968) and Merton (1974), respectively. The Conditional Inference Tree Hothorn et al. (2006), that we apply is a simple variant of the Classification and Regression Tree (CART) algorithm (Breiman et al., 1984) based on a significance test procedure that avoids a bias towards categorical predictors (see Odén, Wedel, et al. (1975)). Conditional Inference Trees are binary trees constructed by splitting a node into two child nodes recursively. Starting from a a root node including the whole learning sample, it proceeds by splitting until some final nodes called leaves. The possible regularizers, $R\big(f(\cdot)\big)$, for any binary tree algorithm are multiple (e.g., depth of the tree, number of nodes, minimal leaf size, information gain at each split) and they are all aimed at penalizing deeper trees. Deeper, more complex, trees need to be regularized as they tend to overfit the training data. Less deep, hence less complex, trees show a lower performance on the training sample, but they are more generalizable. The optimal level of complexity is usually chosen by cross-validation. In a nutshell, the idea behind cross-validation is to train and test the algorithm on multiple independent subsamples in order to choose the level of complexity associated with the best performance on the training sample (out-of-sample performance).

The Random Forest is yet another ensemble method that aggregates different trees to get to a stronger predictive power. Each tree is constructed by randomly picking different variables among all the possible predictors and randomly selecting a subset of the total number of observations (see also Breiman (2001)). Random Forests are linear combinations of multiple and fully grown CARTs. The regularization parameters of the Random Forest include the number of trees, the level of complexity, the number of variables and the observations used to build each tree. Finally, we compare with the Super Learner (Van der Laan et al., 2007), whose intuition is to combine some candidate algorithms, called learners, to create a new prediction algorithm that shall outperform the single learners as it is based on a weighted combination of single-learners' predictions. Van der Laan et al. (2007) show how the estimated loss function of the super learner is at least as large as the best basic learner. For our purposes, the Super Learner we build is a convex combination of a logistic regression model, a Conditional Interference Tree and a Random Forest. Hence, the regularization parameters for the Super Learner are both the ensemble weights and the individual regularization parameters (hyper-parameters) of both the Conditional Inference Tree and the Random Forest.

1.3 Sensitivity checks

We perform a sensitivity analysis for the baseline predictions obtained from the BART-MIA algorithm based on the methodology introduced in Chapter 3 and developed by Bargagli-Stoffi, De Beckker, et al. (2020). We want to make sure that predictions are not sensitive to the inclusion of an unobserved predictor that we can assume is orthogonal to already used predictors but possibly highly correlated with the outcome, i.e. the firms' failures. For our scope, we can generate a synthetic predictor, i.e. a possible confounder Q_i, whose correlation with the outcome is as high as the one of the best observed predictor, but not significantly correlated with any of the other predictors. Hence, we can check how its inclusion affects unit level predictions. If the differences in the unit level predictions between the original model and the augmented model are not significant, then the original model already captures most of the signal in the data. Results are reported in Section 4.

2 A case for *'zombie firms'*

Apparently, the problem of assessing which firms are in trouble could seem trivial because, according to basic economic theory, competitive markets should do their work and let non-profitable firms succumb. Therefore, in absence of market failures, non-viable firms are simply the ones that go bankrupt. Originally, the notion of *zombie firms* related to the phenomenon of *zombie lending*, when banks keep credit flowing to otherwise insolvent borrowers. Sometimes, *zombie lending* has been a deliberate strategy to avoid budget restructuring, while only apparently comply with capital standards set by financial regulations, as it was the case of Japanese banks in the 1990s (Caballero et al., 2008; Peek & Rosengren, 2005). More recently, Schivardi et al. (2017) study the Italian firms during the financial crisis and they find that credit misallocation after *zombie lending* increases the failure rate of healthy firms while reducing the failure rate of non-viable firms. The reason is that under-capitalized banks can decide to cut credit to more viable projects to avoid a public disclosure of non-performing loans in their portfolio. Paradoxically, in times of financial crises, *zombies* may be relatively more resilient thanks to a continuous access to financial resources.

But what is exactly a *zombie firm*? There is no consensus on the exact meaning of the *zombie* attribute beyond its evocative power. In the absence of a clearer guidance from theory, previous scholars just adopt different working thresholds based on the proxy assessment of one or more available financial indicators. Ideally, one should consider firms' competitiveness and financial constraints in a dynamic perspective, having in mind the entire horizon of future events. After considering all future threats and opportunities, and how they mirror on the flows of firm-level profits and financial resources, one would obtain the theoretical identification of *zombie firms*. In absence of the latter, the empirical identification is left to the creativity of the scholars. The seminal work by Caballero et al. (2008) defines *zombies* as some firms that receive subsidies under the form of bank credit after observing how present interest payments compare to an estimated benchmark of debt structure and market interest rates. On the other hand, McGowan et al. (2018) assume that *zombies* are old firms that have persistent problems meeting their interest payments, although the policy discussion is then centered on the macroeconomic impact of low-productivity firms. Bank of Korea (2013) specifically looks at the interest coverage ratio being lower than one over a span of three years. Bank of England (2013) explicitly overlooks financial management and considers firms that have both negative profits and negative value added, hence focusing on the core activity of a company.

According to us, the direction of previous empirical works is clear: one wants to deduce from actual financial accounts the 'viability' of companies. If they do not seem in good shape now, the firm will likely be in troubles in the next future. After that, one may be interested in estimating the aggregate impact that non-viable firms have on the overall welfare, while others may focus on the specific misallocation of financial resources, as in the case of credit scoring.

From our point of view, the empirical identification of *zombie firms* is a perfect case study for the application of machine learning techniques to firm-level data. In fact, we must acknowledge that any chosen present firm-level indicator, or set thereof, aimed at the identification of *zombies* is by nature imperfect.

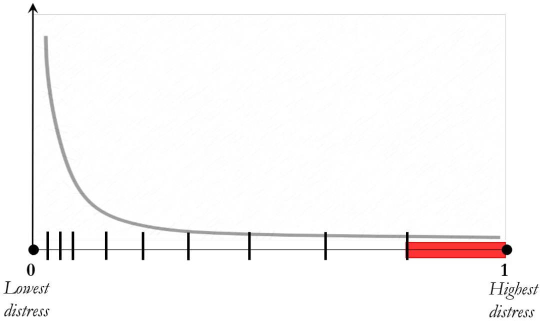
We propose a working definition based on the predictions of firms' failures. Conventionally, we assume that *zombies* are firms that persist in a high risk of failure for three consecutive years. We believe that setting a working threshold of high risk above 90% realistically allows encompassing most difficult situations. That is, we consider the deciles along the predictions $\hat{f}(\mathbf{X}_{it-1})$, where each $q_{j,t}$ is the threshold for the jth decile of failing probability at time t:

$$Q_{j,t} = \begin{cases} 1 & \text{if } \hat{f}(\mathbf{X}_{it-1}) \geq q_{j,t} \cap Y_{it} \neq 1, \\ 0 & \text{otherwise}, \end{cases}$$

where $Y_{it} \neq 1$ indicates that firm i did not fail at time t. Since there is no consensus about the most appropriate threshold to be used, in Appendix 6, we show the results that one would obtain by choosing a less conservative threshold.

As we plot in Figure 34, the idea is that we can assess the distance of each firm from the highest financial distress, which we conventionally may assume in red above the 9th decile. In fact, following analyses will show that the highest decile is the segment of the risk distribution after which it is very difficult to transit back to lower distress.

Figure 34: Fictional distribution of failure's probability

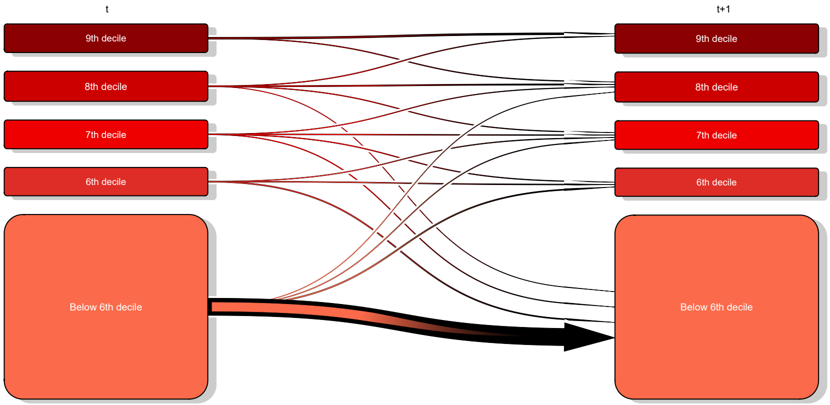


In general, we can say that, based on the information we observe, a firm in good shape does not easily shift into financial distress, but when it does it is difficult to get out of it. Against this evidence, it makes sense for us to set a working threshold at the 9th decile, as we realistically would encompass most difficult situations of the business life. In Table 21, we report a transition matrix for the firms that did not fail, based on elaborations over the entire period of analyses in 2008-2017. In Figure 35, we also visualize a Sankey plot using the same information (Kennedy & Sankey, 1898). We observe that firms that our predictions locate beyond the 9th decile in a representative year t do not have a high chance to improve in $t + 1$. See also the constant trends we report in Figure 36. The bars of the graph indicate the transition of *zombie* firms in the years following predictions: i) to failure (bright red); ii) to a relatively lower distress (dark grey); iii) to a permanence in a *zombie* status. Negligible and non-visible is the share of firms that transit to an area of no distress (bright grey), below the median prediction. Please note that we cannot report year 2017, as we cannot compare with actual observations in fol-

Table 21: Transitions across deciles of risk

t / $t+1$	9th decile $t+1$	8th decile $t+1$	7th decile $t+1$	6th decile $t+1$	Below 6th decile $t+1$	Total $t+1$
9th decile t	0.64	0.24	0.07	0.02	0.03	1.00
8th decile t	0.22	0.39	0.21	0.06	0.12	1.00
7th decile t	0.08	0.20	0.28	0.23	0.21	1.00
6th decile t	0.04	0.08	0.23	0.30	0.35	1.00
Below 6th decile t	0.01	0.02	0.03	0.06	0.88	1.00

Figure 35: Transitions across deciles of risk: a visualization



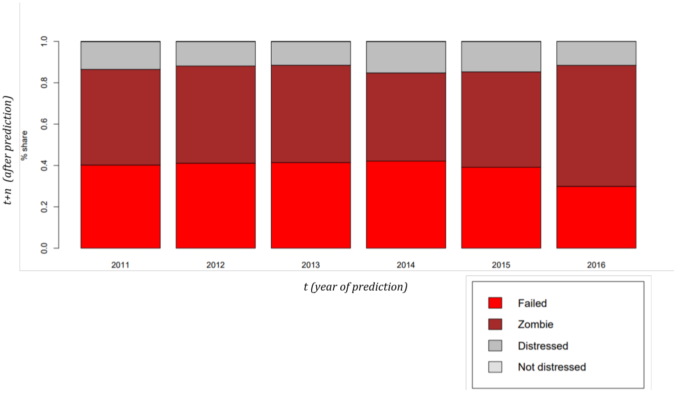
lowing years. The greatest majority of them (64%) gets stuck in the same highest-risk category, 24% transit to the 8th decile, and only 12% are able to recover and reach a more reasonable level of financial distress. Interestingly, the 9th decile is quite difficult to reach from the bottom of the distribution, as only 8%, 4% and 1% of firms that are below the 8th, 7th and 6th decile, respectively, are observed to transit to a situation of highest distress. Moreover, about 88% of company stay below the 6th decile in the entire period of analysis.

We believe that a period of at least three years allows discounting some firms' break-even strategies, hence reducing the share of false positives in the zombie status. Hence, we define *zombie firms* those firms that persist for at least three years beyond the 9th decile of the risk distributions:

$$1(\sum_{t=1}^{3} Q_{j,t} = 3) \tag{4.3}$$

where the risk distributions are the ones that we estimate based on the BART-MIA algorithm. We argue that if a firm is just temporarily under distress, because it is waiting for a break-even after some new investment choices, we should however observe a progressive decrease in the estimated probability of failure over a relatively short span of time, when financial accounts start improving. For a similar reason, we propose to exclude from the perimeter of *zombies* all start-ups for three years, i.e. considering they need some time to reach their break-even.

Figure 36: Transitions after predictions of a *zombie* status



Of course, ours is just a working definition that can be refined in different contexts, while moving along the predictions of failures. In general, however, we argue that our framework improves on previous attempts to define the viability of a firm based on single indicators, because it reduces the error margin while providing an entire ranking of possibilities, given our probabilistic measure, rather than an in-or-out binary forecast.

2.1 Predictors of zombies

In this section, we propose to perform a LOGIT-LASSO (Ahrens et al., 2019) to spot a list of best predictors in a *zombie* status. The functional form in a panel setting is the following:

$$_{\beta\in\mathbb{R}^p} \ \frac{1}{2N}\sum_{i=1}^{N}\left(y_{it}(x_{it-1}^{T}\beta) - log(1+e^{(x_{it-1}^{T}\beta)})\right)^2 \quad \text{subject to } \|\beta\|_1 \leq k \qquad (4.4)$$

where y_{it} is a binary variable equal to one if a firm i had been fallen above the 9th decile of high risk for at least three years until it is observed at time t, while it is zero otherwise. Any x_{it-1} is a lagged predictor chosen in $\mathbb{R}^p$ at time $t-1$, whereas $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$ and $k > 0$. The constraint $\|\beta\|_1 \leq k$ limits the complexity of the model to avoid overfitting, and k is chosen, following Ahrens et al. (2019), as the value that maximises the Extended Bayesian Information Criteria (Chen & Chen, 2008). To account for the potential presence of heteroskedastic, non-Gaussian and cluster-dependent errors we adopt the rigorous penalization originally introduced by Belloni, Chernozhukov, and Wei (2016).

Finally, we run a logistic regression with year and sector fixed effects to assess the statistical association between the set of best predictors selected in the previous LOGIT-LASSO and the *zombie* status:

$$\begin{aligned} \hat{\beta}_{post} &= \ _{\beta\in\mathbb{R}^p} \ \frac{1}{2N}\sum_{i=1}^{N}\Bigg(y_{it}(\sum_{t=1}^{T+1}\alpha_t B_{it} + \sum_{s=1}^{S}\delta_s W_{sit} + x_{it-1}^{T}\beta) \\ &- \ log(1+e^{(\sum_{t=1}^{T+1}\alpha_t B_{it}+\sum_{s=1}^{S}\delta_s W_{sit}+x_{it-1}^{T}\beta)})\Bigg)^2 \qquad (4.5) \end{aligned}$$

$$\text{subject to } \beta_j = 0 \text{ if } \tilde{\beta}_j = 0,$$

where $\tilde{\beta}_j$ is a predictor from (4.4), B_{it} and W_{sit} are sets of time and sector dummy variables that allow controlling for variables that are constant across entities but vary over time, on one hand, and for factors that differ across sectors but are constant over time, on the other hand. We emphasize once again how our entire strategy, including the latter stage, is not intended to identify drivers of failures, since ours is a pure prediction problem. Yet, the identification of a list of best predictors allows us underlying how different indicators may contribute with different pieces of information. Eventually, the list of predictors can change over time and in different economic environments, as we will show in the case of Italy.

3 Data and preliminary evidence

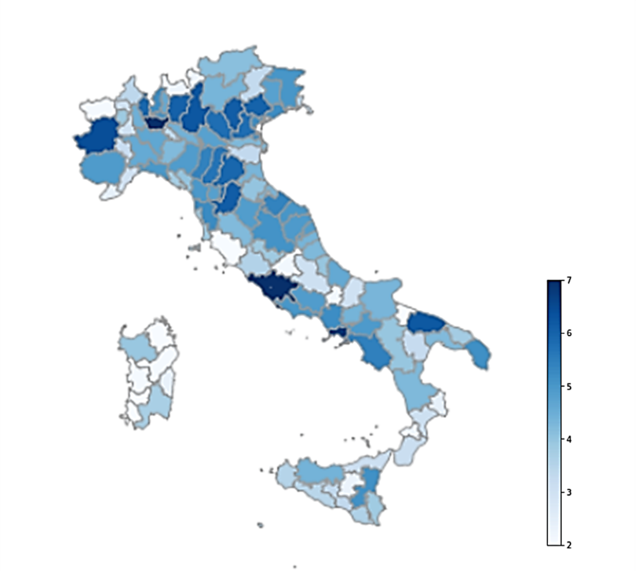
We source financial accounts from the Orbis database[5], compiled by Bureau Van Dijk, for manufacturing firms active in Italy at least one year in the period 2008-2017. First of all, we exploit two main variables that help identifying firms' failures: the status of a firm and the date in which it becomes inactive. Table 22 depicts our sample coverage by status in the period of analysis. We assume that a company failed in the first year when it is reported as being "Bankrupted", "Dissolved", or "In Liquidation". Altogether, the share of exiting firms in our sample constitutes about 5.7% of the entire sample, which is a figure not too far from the average official figure (6.3%) indicated by ISTAT, the national statistics office, in the same period. Figure 37 maps the firms' failures by NUTS 3-digit regions in logarithmic scale. As expected, we find a concentration of failures in metropolitan regions but full representativeness of the entire Italian territory, as we detect failures in any region in our period of analysis.

Table 22: Firms by status

Status	Active	Bankrupted	Dissolved	In Liquidation	Total
Sample	287,586	1,533	8,540	7,221	304,906
Percentage	94.33%	0.50%	2.80%	2.37%	100%

[5] Orbis firm-level data have become a common source for world financial accounts. See for example, Gopinath et al. (2017) and Gal (2013). Coverage of smaller firms and some balance sheet information can change from country to country (Kalemli-Ozcan et al., 2015), according to filing requirements by national business registries. Moreover, disclosure of specific financial accounts can change according to different national regulations.

Figure 37: Geographic Coverage



Note: all regions are covered in our sample.

Additionally, we exploit a battery of economic and financial indicators to train our predictive models. The battery includes: i) original firm-level financial accounts; ii) indicators that have been adopted in previous literature to proxy firm-level financial constraints; iii) indicators that have been used so far to spot *zombie firms*; iv) indicators that are adopted as an alert on firms' crises in the recent Italian bankruptcy law[6]. Any predictor we consider is described in detail in Table 23.

Please note how many of the indicators that we pick as predictors have been used in very different frameworks to assess how much a firm is in trouble. From our perspective, they cannot be interpreted as a drivers of failure. It is enough if they contribute with an albeit small piece

[6] A recent reform of the bankruptcy law (L. 155/2017 and DL 14/2019) proposes an early warning system based on indicators proposed by practitioners. The purpose is to spot firms that are in trouble before it is too late, preserving entrepreneurial abilities and finding a way out of the crisis.

Table 23: Panel (A): List of predictors for firms' failures.

Variables	**Description**
Operating Revenues, Material Costs, Costs of Employees, Added Value, Taxation, Tax and Pensions' Payables, Financial Revenues, Financial Expenses, Interest Payments, Number of Employees, Net Income, Cash Flow, EBITDA (Earnings before interest, Taxation, Depreciation and Amortization), Total Assets, Fixed Assets, Current Assets, Shareholders' Funds, Retained Earnings, Long-Term Debt, Loans, Current liabilities	Original firm-level financial accounts expressed in euro.
Corporate control (Control)	A binary variable equal to one if a firm belongs to a corporate group.
Number of patents	The stock of patents granted to a firm by a patent office.
Number of trademarks	The total number of trademarks issued to the firm by national or international trademark offices
Consolidated accounts	A binary variable equal to one if the firm consolidates accounts of its subsidiaries
NACE rev. 2	A 4-digit industry of affiliations for each firm, following European classification NACE rev. 2.
Liquidity ratio	(Current assets - Stocks) / Current liabilities
Solvency ratio	(Shareholders funds / (Non current liabilities + Current liabilities)) * 100
Capital Intensity (CI)	It is a ratio between the fixed assets and the number of employees.
Labour productivity	It is a ratio of added value over the number of employees.
Enterprise Value (Listed companies only)	It is a synthetic value sourced from the original Orbis database for each listed company, and it is calculated considering other 10 comparable listed companies in terms of Market Capitalization, Minority Interest, Preferred shares, Long Term debt, Loans, Other short term debt, Cash.
Interest Coverage Ratio (ICR)	It is calculated following the forumula $ICR = \frac{EBIT}{Interest\ Expenses}$. It considered an indicator of *zombies* when it is less than one for three consecutive years by Bank of Korea (2013) and McGowan et al. (2018).

Panel (B): List of predictors for firms' failures.

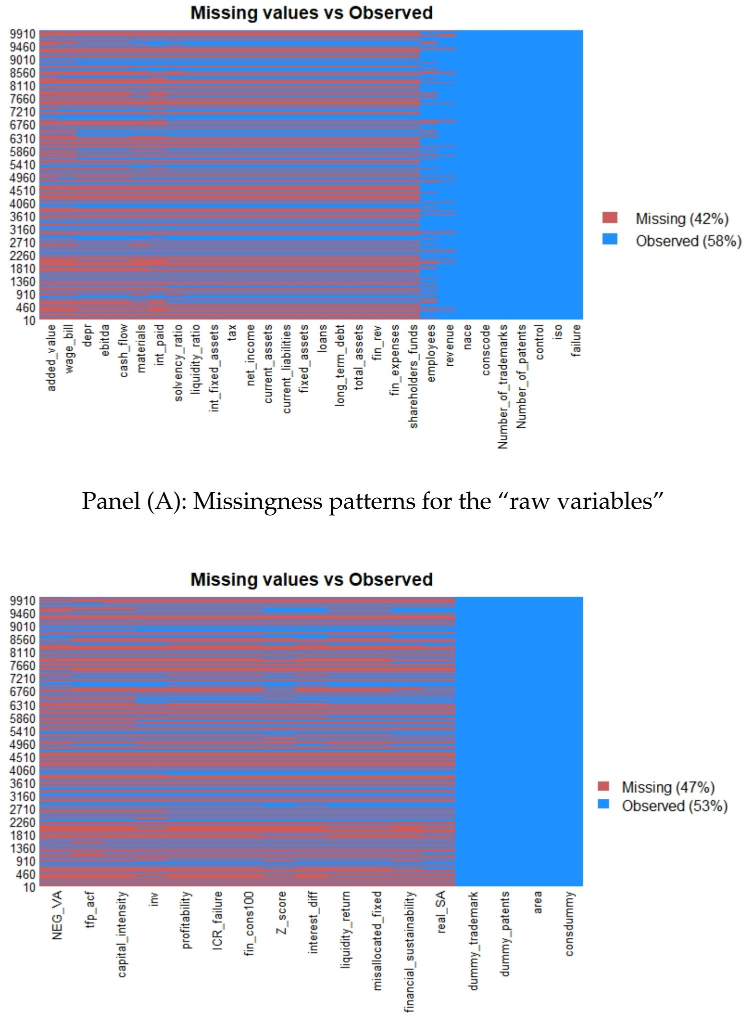
Variables	**Description**
Financial Misallocation (FM)	It is a binary indicator adopted by Schivardi, Sette and Tabellini (2017) for catching *zombie lending*, based on both $ROA \frac{\frac{1}{3}\sum_{t=1}^{3} EBITDA_t}{Total\ Assets} < prime$ and $Leverage = \frac{Financial\ Debt}{Total\ Assets} > \tilde{L}$, where $prime$ is the measure of the cost of capital for firms with Z-score equal to 1 or 2, and where $\tilde{L}$ is the median value of leverage in the current year for firms that exited in two following years.
TFP	It is the Total factor productivity of a firm computed as in Ackerberg et al. (2015)
Financial Constraint Indicator (FCI)	It is a proxy of financial constraints as in Nickell and Nicolitsas (1999), calculated as a ratio between interest payments and cash flow
Productive Capacity	it is an indicator of investment in productive capacity computed as (fixed assets in t/ (fixed assets in t-1 + depreciation in t-1))
Negative Added Value	It is a binary variable adopted by Bank of Korea (2013) for zombie firms equal to one when added value is less than zero.
Size-Age	It is a synthetic indicator proposed by Hadlock and Pierce (2010), equal to $-0.737 * \log(totalassets) + 0.043 * \log(totalassets))^2 - 0.040 * age$.
Profitability	It is the Profitability indicator by Schivardi et al. (2017)
Financial sustainability	It is a ratio of financial expenses over revenues
Capital Adequacy Ratio	It is the ratio of shareholders' funds over the sum of short and long term debt
Liquidity Returns (LR)	It is the ratio of cash flow over total assets
Tax and Pension Payables	It is the ratio of the sum of tax and pension payables over total assets.
Interest Benchmarking (BID)	It is a *zombie* proxy proposed by Caballero et al. (2008) calculated as $R^* = rs_{t-1}BS_{it-1} + (\frac{1}{5}\sum_{j=1}^{5} rl_{t-j})BL_{it-1} + rcb_{5y,t} \cdot Bonds_{it-1}$, where BS_{it-1} are short-term bank loans, BL_{it-1} are long-term bank loans, rs_{t-1} are the average short-term prime rate in year t, rl_{t-j} are the average long-term prime rate in year t, $Bonds$ are the total outstanding bonds, $rcb_{5y,t}$ is the minimum observed rate on any convertible corporate bond issued over the previous 5 years.

of information on the healthiness of a firm. In our prediction framework, they could even border on multicollinearity, as for example in the case of different measures of efficiency, when we include both Total Factor Productivity and labor productivity. As far as we are not interested in the identification of the single contribution made by each predictor to the outcome, multicollinearity will not affect the ability of a model to predict the outcome (Makridakis et al., 2008; Shmueli, 2010; Vaughan & Berry, 2005). Interestingly, in Figure 38 we visualize a map of the ensemble of the predictors to bring out patterns of missing values in brighter colors. We register that some predictors are relatively more missing than others. After we run a series of chi-squared tests, we do find that there is a statistical association between observed patterns of missing values and the event of a firm failure. The results from these test are reported in Table 22.

For our analysis, the main issue is sample selection, when observations are selectively missing for some categories of firms, potentially introducing a bias on the identification of non-viable firms. Obviously, smaller firms are often exempted from full financial reporting, especially when they are not quoted at the stock exchange. In fact, smaller firms may also be the ones that suffer relatively more from financial constraints, hence they are more likely distressed in times of crisis. On the other hand, firms of any size more likely avoid disclosing negative outcomes when in distress. The application of a BART-MIA procedure, as described in Kapelner and Bleich (2015) allows us considering the patterns of missing values as a further predictor of the outcome in following analyses.

The missingness analyses reported above are particularly relevant for the scope of this Chapter. The main issue is sample selection when observations are selectively missing for some categories of firms. In this case, we have two potential sources of sample selection bias: i) firms in distress *vis à vis* firms not in distress, because the first can have an incentive to disclose less information than the latter; ii) smaller firms *vis à vis* bigger firms because the first are often exempted from a full financial report, ac-

Figure 38: Missing values from a sample of 10,000 observations in the period 2008-2016



Panel (A): Missingness patterns for the “raw variables”

Panel (B): Missingness patterns for the indicators built using the “raw variables”.

Table 22: Missingness tests by status

	Firm's failure		Test Statistic
	0	1	
	$N = 287587$	$N = 17319$	
BID : 0	38% (110524)	61% (10530)	χ^2_1=3414.25, P<0.001
BID : 1	62% (177063)	39% (6789)	
ICR : 0	37% (105907)	49% (8422)	χ^2_1=970.93, P<0.001
ICR : 1	63% (181680)	51% (8897)	
Negative AV : 0	34% (98014)	63% (10915)	χ^2_1=5958.81, P<0.001
Negative AV : 1	66% (189573)	37% (6404)	
FCI : 0	37% (105904)	49% (8419)	χ^2_1=968.27, P<0.001
FCI : 1	63% (181683)	51% (8900)	
FM : 0	39% (112560)	54% (9276)	χ^2_1=1415.82, P<0.001
FM : 1	61% (175027)	46% (8043)	
TFP : 0	36% (104345)	38% (6600)	χ^2_1=23.52, P<0.001
TFP : 1	64% (183242)	62% (10719)	
Solvency : 0	41% (118851)	63% (10897)	χ^2_1=3115.5, P<0.001
Solvency : 1	59% (168736)	37% (6422)	
Liquidity : 0	42% (119357)	72% (12543)	χ^2_1=6362.72, P<0.001
Liquidity : 1	58% (168230)	28% (4776)	
Size-age : 0	42% (120260)	75% (12989)	χ^2_1=7310.19, P<0.001
Size-age : 1	58% (167327)	25% (4330)	
LR : 0	39% (112561)	54% (9277)	χ^2_1=1416.88, P<0.001
LR : 1	61% (175026)	46% (8042)	
Labour Product. : 0	34% (97253)	36% (6221)	χ^2_1=32.23, P<0.001
Labour Product. : 1	66% (190334)	64% (11098)	
Profitability : 0	37% (105907)	49% (8422)	χ^2_1=970.93, P<0.001
Profitability : 1	63% (181680)	51% (8897)	
Fin. sustainability : 0	41% (117294)	64% (11119)	χ^2_1=3673.94, P<0.001
Fin. sustainability : 1	59% (170293)	36% (6200)	
Capital intensity : 0	36% (104122)	42% (7325)	χ^2_1=261.17, P<0.001
Capital intensity : 1	64% (183465)	58% (9994)	

Note: numbers after percents are frequencies. Test used: Pearson Chi-squared test. Please find the entire battery of test in the Online appendix.

cording to Italian regulation [7]. In fact, the two sources can overlap, since smaller firms may also be the ones that suffer relatively more from financial distress. The application of a BART-MIA procedure, as described in Section 1, allows us considering both sources of sample selection when patterns of missing financial accounts emerge because the algorithm includes such patterns as a further predictor of firms' failures.

Interestingly, as reported from panel (B) of Figure 38, the most missing variables are the ones that have been used in previous works as proxies for *zombies*. Take for example the case of the Interest Coverage Ratio (ICR), derived as a ratio between the earnings before interest and taxes (EBIT) and the interest expenses. When ICR is lower than one, Bank of England (2013) assumes that a firm is a *zombie* since it has problems in meeting financial obligations. In our sample, we find that about 19% of firms have an ICR smaller than one but, at the same time, there are 62.50% firms whose information on ICR is not present at all. Further, according to (Bank of Korea, 2013), a negative value added is the most appropriate proxy for assessing a *zombie* status, since in this case the value of intermediate inputs is higher than the value of the output realized on the market. In the case of Italy, about 64.27% of companies do not report any value in the sample, while about 3% of them reports a negative figure. Clearly, negative value added is a stricter condition than negative profits, as a firm can have no profits without destroying economic value. Indeed, firms' profitability is at the core of two very similar proxies of *zombies* adopted by Schivardi et al. (2017), when they compare firm-level profits with a fixed benchmark. Reproducing the same exercises, we would find about 3% of firms are in trouble, while a large amount of firms (62.50%) does no report any information at all. Finally, Caballero et al. (2008) and McGowan et al. (2018) adopt an interest benchmark to compare the interest paid by a firm to obtain financial resources

[7] Following Italian civil law, companies that do not quote financial activities on the stock exchange have the possibility to present more aggregate financial accounts when their size does not exceed simultaneously two of the following thresholds, for one or two consecutive periods: i) 4,400,000 euro of total assets; ii) 8,800,000 euro of operating revenues; iii) 50 employees. A limited financial statement always includes main items at the first or second digit of aggregation.

and some reference on the cost opportunity to invest the resources in alternative safer assets. For example, when we take as benchmark the yields on state-issued bonds with a maturity of ten years (European Central Bank), we find that up to 42% of sample firms do pay higher interests, although 60.29% of firms do not report any information. Finally, we include as a predictor of a firm's failure also an estimate of Total Factor Productivity (TFP), following the methodology proposed by Ackerberg et al. (2015), which takes into account the simultaneity bias deriving from *ex-post* adjustments in the combination of factors of production. In this regard, firm-level TFPs allow us considering the ability to transform factors of production and sell on the market. We may assume that less productive firms are also the ones that may encounter more financial constraints.

4 Results

In this Section, we report the results of our analysis. Subsection 4.1 depicts and discusses the results of the proposed machine learning methodology as compared to (i) standard econometric techniques, (ii) other state-of-the-art machine learning algorithms and, (iii) financial indicators such as Z-score Altman (1968) and Distance-to-Default Merton (1974). Finally, subsection 4.2 displays the results for the analysis of *zombie* measures which have been used so far and their validation.

4.1 The performance of our preferred ML method

Since our *zombie* indicator is built on the ability of the algorithm to properly predict firm level probability of failures, it is central to validate different algorithms and indicators performance in this task. Here, we compare the predictions of different techniques on which firms should or should not fail with the observed firms statuses.

Before getting in the nuts and bolts of the performance of the proposed BART-MIA algorithm (and how it compares to other methodologies used to assess firm's financial distress), let us briefly discuss a peculiarity of our data: the unbalanced distribution of the outcome. Indeed, as shown in 22, the largest part of the firms (94.3%) do not exit the market in the time span of our analysis (zero inflated distribution). What is the problem with dealing with such an unbalanced setting? Take as an example a methodology that would predict all the firms in our dataset to be failed. Obviously, such a methodology would be totally uninformative, but it still would get an accuracy measure (correctly classified observations over total observations) of 94.3%.

In order to deal with unbalanced output variables, different methods have been proposed. The two that are most widely used by scholars and practitioners are either *under-sampling* observations for the label that accounts for the largest majority of observations, or *over-sampling* observations for the other label (Gruszczyński, 2019; L. Zhou, 2013). We argue that both these methodology would be sub-optimal in our setting. Indeed, as we deal with a scenario where the missingness pattern in the predictors are potentially informative, under-sampling the failed firms would artificially induce a higher association between the missing observations and the outcome, while over-sampling would induce the opposite effect. Hence, we do not use these methodology, but we accommodate for the unbalanced outcome data by using performance measures that account for such settings. In this way, we use all the information in our dataset without artificially introducing any bias in the data. In particular, the main goodness-of-fit measures that we employ are the F1-score (Van Rijsbergen, 1979), the Balanced ACCuracy (BACC) and the area under the Precision-Recall curve (PR)[8]. These three goodness-of-fit measures account, in different ways, for unbalanced outcomes. In addition, we compare them with performance measures that do not adjust for unbalanced outcomes such as the Area Under the receiver operating characteristic Curve (AUC, Hanley and McNeil (1982)) and the adjusted R^2.

Table 23 depicts the comparative results of BART-MIA, logistic regression (logit), conditional inference tree (Ctree), random forest and super learner. The results are obtained through a standard two folds cross-

[8] For a detailed description of these performance measures we refer to Fawcett (2006).

Table 23: Machine learning techniques and out-of-sample prediction accuracy

	Models' Horse Race						
Model	AUC	PR	F1-score	BACC	R^2	Train Obs	Test Obs
Logit	0.8896	0.3576	0.2098	0.8433	0.0829	83,537	9,282
Ctree	0.8889	0.3568	0.2000	0.7804	0.0654	83,537	9,282
Random Forest	0.9050	0.4262	0.2257	0.8515	0.0922	83,537	9,282
Super Learner	0.9073	0.4311	0.2232	0.8666	0.0945	83,537	9,282
BART-MIA	0.9667	0.7484	0.6328	0.8993	0.4038	83,537	9,282

validation procedure that assigns 90% of the observations to the training sample and 10% to the test sample (Devijver & Kittler, 1982). Please note how BART-MIA could always use a larger training sample, including firms that report missing values on predictors. This is an additional advantage of our baseline methodology. However, for sake of comparison, all methodologies in Table 23 have been trained on the same subset of roughly 84,000 observations for the training and 9,000 for the testing, without considering missing values[9].

BART-MIA outperforms a standard econometric technique such as logistic regression, and all the other machine learning techniques employed for the analysis[10]. In particular, the increase in the performance of BART-MIA as compared to the super learner[11] ranges between 4%

[9] As a robustness check, we implemented an alternative methodology in which we preprocessed a subsample of data to impute missing values using a CART-based methodology (Buuren & Groothuis-Oudshoorn, 2010) similar to the one proposed by T. K. White et al. (2018). Then, we applied methods alternative to BART-MIA. Results are in line with the ones reported in the Chapter and with previous predictive analysis on ORBIS data for Italian enterprises by Weinblat (2018). We don't find a sensitive improvement in the prediction performance of the alternative methods. For additional details on this robustness analysis we refer to the online Appendix.

[10] Let us note that, in our scenario, a fine hyper-parameters tuning for the machine learning techniques that we are comparing to BART-MIA does not lead to any sizable increase in the performance. This hints at the fact that any fine hyper-parameters tuning would overturn the gap in performance between BART-MIA and all the other machine learning methodologies.

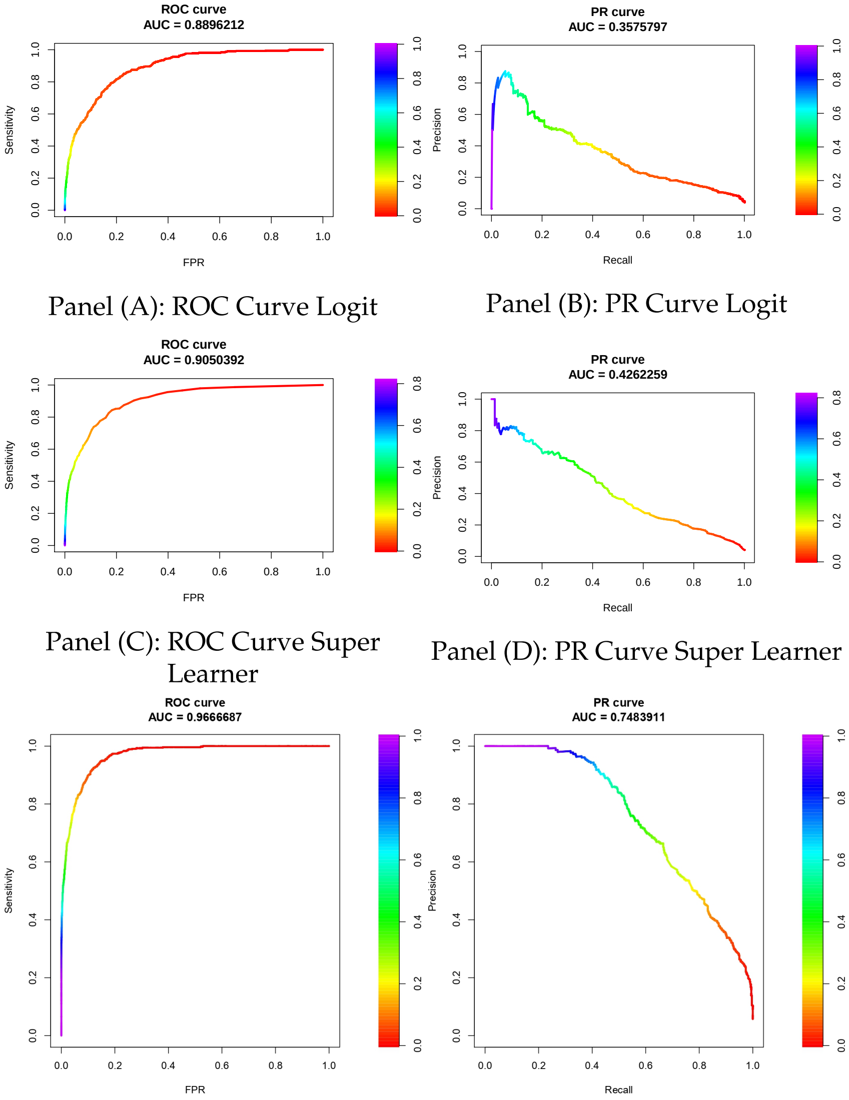
[11] It is interesting to highlight that the super learner is performing better than the random forest with respect to the accuracy, but not with respect to the F1-score. This is due to the fact that the super learner algorithm (Van der Laan et al., 2007) is optimized in a way to find a convex combination of learners that minimizes the accuracy of the ensemble method. In this particular case, this strategy is not optimal as the output data are unbalanced.

and 327%, depending on the various performance measures. These results show that the additional information coming from the missingness patterns in the data are particularly valuable. In particular, Figure 39 provides an insight on how BART-MIA improves its performance with respect to the other techniques. On the left column, we report the results for the AUC performance measures of logit, super learner and BART-MIA respectively. On the right column, we report the results for the PR curve for the same techniques. Both AUC and PR are performance measure that vary between 0 and 1, where 0 indicates complete misclassification, 1 perfect prediction. For AUC values around 0.5 indicate that the model has the same predictive power as a random assignment. On the one side, AUC takes into account the *sensitivity/recall* (true positive rate) and the *false discovery rate* (how many observations where misclassified as positives). On the other side, PR accounts for sensitivity/recall (true positive rate) and for *precision* (positive predicted values: the proportions of positive results that are true positive). Hence, while the performance with respect to both sensitivity/recall and false positive rate are similar (even though BART-MIA performs better than the other models), the real boost in the performance comes from a higher precision (one can check this by comparing Panels (B), (D) and (F) in Figure 39). This means that BART-MIA has a higher ability to precisely predict firms that will fail one year from now. This is central for our analysis, as our *zombie* indicator is directly built on the ability of machine learning to correctly assess the failure probability, assigning distressed firms to higher probability of default.

Moreover, BART-MIA shows a higher performance than other machine learning algorithms employed for the same task on similar data such as the ones proposed by Moscatelli et al. (2019) using Italian data with additional unstructured information, and Weinblat (2018) using the same ORBIS data. For a detailed analysis on how the current method would compare with state-of-the-art machine learning methodologies we refer to Bargagli-Stoffi, Niederreiter, et al. (2020).

On top of that, BART shows a high stability in its prediction as shown by the *sensitivity of prediction analysis*. The original model is stable as the

Figure 39: Out-of-sample goodness-of-fit



Panel (A): ROC Curve Logit

Panel (B): PR Curve Logit

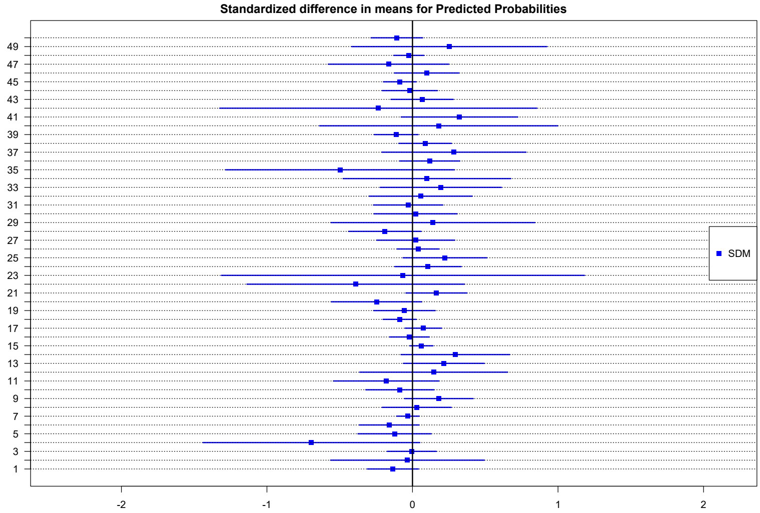
Panel (C): ROC Curve Super Learner

Panel (D): PR Curve Super Learner

Panel (E): ROC Curve BART-MIA

Panel (F): PR Curve BART-MIA

Figure 40: Standardized differences in means



addition of a new highly explanatory predictor is not affecting in a sizable way the predicted values. Figure 40, shows the standardized differences in means between the prediction of the original BART model and the predictions of the "augmented" BART model for 50 sampled units. One can easily see that there is no evidence of significant differences between the unit level predictions. Indeed, we find that just a negligible percentage of unit level predictions are different as the correlation between the confounder and the outcome reaches 0.5 (higher than the best predictor in the model). For lower correlation values, the differences are even more negligible. Such an analysis informs us that the predictive model that we implement is capturing enough signal from the data.

Thus far, we compared different econometric and machine learning techniques with respect to their ability of correctly discriminating between viable and non-viable firms. Table 24 depicts the results for precision and false discovery rate of two proxy models widely used to assess the viability of firm's investment projects: the *Z-Score* (Altman, 1968) and the *distance-to-default* (Merton, 1974). The results are obtained setting a range of thresholds ranking from the 10th to the 1st percentile, denoting increasingly higher firm's distress. In order to make these results comparable to the ones of BART-MIA the performance measures are computed just on the set of observations used for the test sample[12]. Distance-to-default shows both a higher precision (0.2680 vs 0.1613) and a lower false discovery rate (0.7320 vs 0.8387) as compared to the Z-score, hinting at a higher performance of this proxy model. However, both models are outperformed by BART-MIA whose precision and false discovery rate are 0.8278 and 0.1722, respectively.

Hence, the proposed BART-MIA methodology outperforms standard econometric, proxy and machine learning models in capturing firms' financial distress. This higher ability to discriminate between viable and non-viable firms, in particular with respect to firm's with high failure risk, is a strength for our definition of *zombies* as surviving non-startup enterprises at high-risk of failure for three consecutive years.

Table 24: Precision and false discovery rate (FDR) of DtD and Z-score

Percentile	Precision DtD	FDR DtD	Precision Z-score	FDR Z-score
1st	0.2680	0.7320	0.1613	0.8387
2nd	0.2680	0.7320	0.1505	0.8495
3rd	0.2680	0.7320	0.1505	0.8495
4th	0.2258	0.7742	0.1371	0.8629
5th	0.2022	0.7978	0.1269	0.8731
6th	0.1759	0.8241	0.1185	0.8815
7th	0.1569	0.8431	0.1108	0.8892
8th	0.1467	0.8533	0.1036	0.8964
9th	0.1411	0.8589	0.0981	0.9019
10th	0.1313	0.8687	0.0969	0.9031

Note: these results should be compared with the ones of BART-MIA. The precision for BART-MIA is 0.8278 (best result for precision is 1) and the false discovery rate is 0.1722 (best result for the false discovery rate is 0). Moreover the analyses are performed on the test set.

[12] Comparing the performance on the training test, could artificially boost the performance of BART-MIA.

4.2 Which are the best *zombie* indicators?

In Section 2.1, we developed a method to assess which are the best financial indicators to catch the *zombie* phenomenon among the ones in the literature (Bank of England, 2013; Bank of Korea, 2013; Caballero et al., 2008; McGowan et al., 2018; Schivardi et al., 2017). In order to do so we proposed a LASSO-based technique. Indeed, after constructing our *zombie* indicator, we employ a LOGIT-LASSO regression to select the variables that best predict a firm being *zombie* and then we use the selected variables for a post-logistic regression.

Table 25: Ranking of predictors by year (*zombies* 9th decile)

Rank	2017	2016	2015	2014	2013	2012	2011
1	FCI*	Control*	Control*	FCI*	FCI*	FCI*	FCI*
2	Control*	FCI*	FCI*	Control*	Control*	LR*	Control*
3	ICR*	Negative AV*	LR*	ICR*	LR*	Control*	LR*
4	Negative AV*	Size-Age*	Negative AV*	Negative AV*	Negative AV*	Profitability*	Negative AV*
5	Size-Age*	LR*	Size-Age*	Size-Age*	Profitability*	Negative AV*	Profitability*
6	LR	ICR*	Profitability*	Profitability	Size-Age*	Solvency*	Solvency*
7	Solvency*	Solvency*	ICR*	Solvency*	ICR*	Size-Age*	Size-Age*
8	BID	Profitability*	Solvency*	LR*	Solvency*	ICR*	ICR*
9	Profitability	BID*	BID	BID	TFP	TFP	Trademark*
10	CI	CI	TFP	Innovation	FM	Innovation	Innovation

Note: The ranking refers to the first ten ranked financial predictors by logit-lasso. The ascetics indicates those variables that are selected from the rigorous logit-lasso as well.

In Table 25, we show the results for the best selected predictors by year. The indicators that rank higher and that are more often included are: the *financial constraint indicator* proposed by Nickell and Nicolitsas (1999) (FCI), the dummy variable *control* that assumes value 1 if the firm is owned by a bigger enterprise and 0 otherwise, *interest coverage ratio* (Bank of Korea, 2013), followed by other *zombie* indicators such as *negative value added* (Bank of England, 2013), liquidity return, solvency and *profitability* (Schivardi et al., 2017) and financial constraint indicators such as the *size-age* proposed by Hadlock and Pierce (2010). However, the indicator used by OECD (McGowan et al., 2018) that is reported as *BID* (*Benchmark Interest Difference*) ranks quite low in all the years. This could be due to the fact that the original indicator developed by Caballero et al. (Caballero et al., 2008) is not precisely reconstructable with the ORBIS data[13]. This analysis highlights that, despite the low predictive power of *BID*, all the other indicators used so far in the literature to describe the dynamics of failure in the market of allocation of credit are fairly good predictors of failures in the market selection mechanism. These two dynamics that should be, in principle, very correlated seem to be, indeed, two sides of the same coin. Moreover, there is not evidence of a strong role of firm level productivity on the chance of being a *zombie*. These result are made robust by running the same selection procedure with a *zombie* indicator defined with respect to a less conservative threshold (namely, the 8th decile), and by proposing an alternative selection method. The results for these additional analyses can be find in Appendix 6.

Table 26 depicts the results from the post-logistic regression analysis. We find a negative association between the enterprise being controlled and its likelihood of being *zombie*, while a positive association is found with respect to the financial/*zombie* indicators such as *size-age*, *FCI*, *ICR*, negative added value and *profitability*. These positive associations indicate that as the financial burden increases, the likelihood of the firm being a *zombie* increases as well. On the opposite side, higher solvency and liquidity return are associated with a lower likelihood of the enterprise being a *zombie*. The same results are found when the analysis is performed on the less conservative *zombie* definition (see Appendix 6 for additional details).

Furthermore, it is important to highlight that the average out-of-sample performance of the post-logistic regressions is higher than the performance of the logistic regression implemented with all the predictors. The adjusted R^2 increases by roughly 100% (from 0.0829 to 0.1670) while the F1-score by 115% (from 0.2098 to 0.4524). This is due to the fact that the logistic regression performed with all the predictors overfits the training data, and consequently has a lower out-of-sample performance. This evidence, on one side informs us on the quality of the selection procedure, while on the other it highlights that scholars and practitioners should rely on a selected number of good-performing indicators in order

[13] This limit is present both in our analysis and OECD's paper.

to spot *zombie firms*.

Table 26: Post-logistic regressions with predictors selected with rigorous logit-lasso (*zombies* 9th decile)

	(2017)	(2016)	(2015)	(2014)	(2013)	(2012)	(2011)
Control	-7.394***	-7.937***	-6.061***	-4.610***	-4.331***	-3.806***	-3.597***
	(-10.23)	(-12.06)	(-17.04)	(-28.56)	(-35.19)	(-42.58)	(-40.13)
Size-Age	0.898***	0.967***	1.056***	0.791***	0.623***	0.370***	0.380***
	(13.25)	(22.13)	(26.98)	(15.87)	(17.65)	(15.45)	(17.35)
FCI	0.814***	1.134***	1.111***	1.356***	2.028***	1.793***	2.723***
	(8.51)	(13.69)	(9.97)	(15.30)	(11.04)	(21.27)	(15.33)
ICR	1.304***	1.068***	0.948***	1.372***	0.717***	0.349***	0.349***
	(11.93)	(14.34)	(12.17)	(19.20)	(10.26)	(3.63)	(5.46)
Solvency	-0.0112***	-0.0161***	-0.0126***	-0.0124***	-0.0123***	-0.0189***	-0.0203***
	(-8.42)	(-12.00)	(-11.02)	(-12.60)	(-5.93)	(-14.69)	(-13.50)
Negative AV	1.083***	1.718***	1.580***	1.222***	1.490***	1.240***	1.142***
	(8.96)	(16.67)	(9.04)	(9.32)	(6.17)	(10.04)	(13.84)
Profitability		0.331***	0.470***	0.329***	0.417***	0.435***	0.300***
		(4.25)	(8.62)	(5.37)	(3.34)	(5.18)	(3.91)
LR		0.367	-1.224***	-2.221***	-0.355	-0.724*	-1.115***
		(1.91)	(-4.11)	(-7.61)	(-0.24)	(-2.08)	(-3.61)
BID		0.516***					
		(15.20)					
Trademark							-1.072***
							(-12.04)
Constant	5.847***	6.542***	7.875***	5.160***	3.402***	0.855**	0.733**
	(8.17)	(13.36)	(19.35)	(10.70)	(9.12)	(2.90)	(3.21)
Year FE	(YES)	(YES)	(YES)	(YES)	(YES)	(YES)	(YES)
Sector FE	(YES)	(YES)	(YES)	(YES)	(YES)	(YES)	(YES)
Observations	78720	80371	82084	82607	82119	80124	76273

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

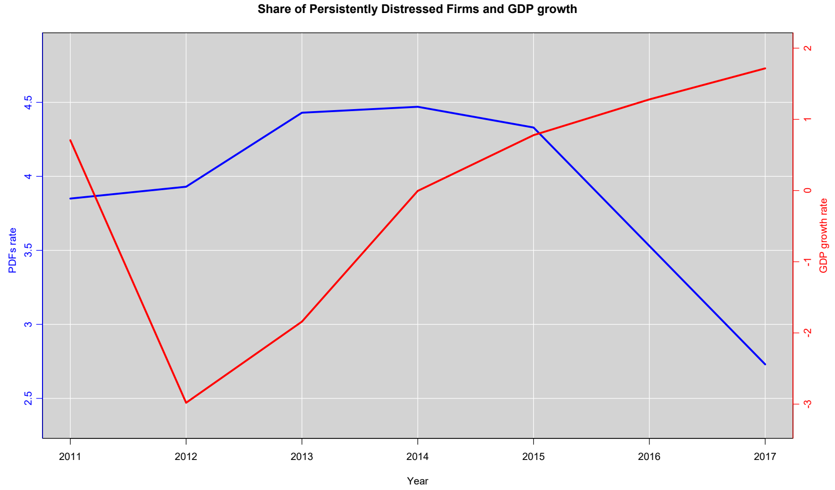
5 Discussion

In this work, we propose machine learning analyses of firm level data to spot non-viable albeit surviving firms, which could hinder reallocation of resources. So called *zombie firms* may be relevant for misallocation of financial resources, on one side, especially during and after financial shocks, as previous references suggest. Also, non-viable low-productivity firms may reduce overall welfare, preventing industrial restructuring. If we take 2011 as a reference year, we can track the amount of added value and employment sunk in *zombie firms* in the following years. Our back-of-envelope estimate is that average employment for firms that were detected as being *zombie*, that failed in the years until 2017, is significantly lower that for non-*zombie* firms (4.40 vs 16.22 employees per firm). The same holds true for the added value (89,853 vs 1,065,850 Euros per year). Hence, we find that *zombies* are smaller firms and their added value accounts just for the 0.34% of the overall added value of the enterprises considered in the analysis (even though *zombies* are roughly 4% of firms).

The implementation of our indicator in the Italian scenario founds a negative lagged correlation between GDP growth and *zombie* firms, rejecting the hypothesis of *zombies* as a cause of the Italian crisis (see Figure 41). Based on this finding, following Ottaviano (2011), we argue that *zombies* can play a role in slowing down the recovery: a bad selection mechanism for firms exiting the market can slow down the recovery during the downswings of the business cycle. Since *zombies* and unemployment seem to follow similar growth patterns, future research should investigate the possibility that for very small, distressed firms, in the presence of a stagnant labour market, survival might be just an alternative to unemployment.

We show how statistical learning can derive non-trivial information on a battery of financial indicator, and successfully identify firms in a persistent high-failure-risk situation. Our preferred machine learning technique, BART-MIA, outperforms both simple proxies and otherwise standard econometric models in predicting firms' failures. This is particularly true in the presence of missing-not-at-random observations where the patterns of missingness would induce a large bias in the construction of standard indicators (i.e., ICR, negative added value, etc.). We argue that the study of *zombies* is central to highlight when the market selection mechanism fails. In the present work, we reconnect the two sides of the *zombie* phenomenon (survival and financial side) into a single probabilistic indicator, overcoming the limitations of deterministic indicator.

Figure 41: *Zombie firms* (9th decile) and GDP



The machine learning based indicator proposed is robust to a potential confunding factor orthogonal to the observed predictors and to variations in the training samples as we show through our novel *sensitivity of prediction* analysis.

Moreover, the method that we implemented for the validation of the current, deterministic indicators highlights how *financial constraint indicators*, such as the ones proposed by Nickell and Nicolitsas (1999) and Hadlock and Pierce (2010), and *zombie* indicators such as *interest coverage ratio* (Bank of England, 2013), *negative added value* (Bank of Korea, 2013), *profitability* (Schivardi et al., 2017), liquidity return and solvency seem to be the most reliable indicators.

Following Calligaris et al. (2016), we believe that a revision of national regulations for bankruptcy and a higher efficiency of the judicial system may be needed to reduce the presence of *zombie firms*. On the other hand, a bigger attention to the phenomenon of *zombie* lending is needed to prevent credit misallocation and crowding-out of financial resources at the expenses of more viable firms with profitable projects of investments. To the best of our knowledge, scant attention has been devoted to the phenomenon of *zombie lending* by the public sector, although regulatory agencies are moving forward to investigate its impact on the banking sector (e.g., Schivardi et al. [2017]).

We argue that our machine learning indicator assumes a particular relevance from a policy perspective as some countries, particularly in Europe (following the European Directive 2012/30/UE), are re-calibrating in their business failure legislation. Some countries are proposing to include in the financial assessment of firms' viability indicators tailored to signal the crises of firms in an early stage (i.e., the Italian Law October 19th 2017, n. 155). In this context, the higher predictive power of machine learning algorithms can boost targeted financing policies that lead to safer allocation of credit either on the extensive margin, reducing the number borrowers by lending money just to the less risky ones, or on the intensive margin (i.e., credit granted), by setting a threshold to the amount of credit risk that banks are willing to accept (Bargagli-Stoffi, Niederreiter, et al., 2020). Indeed, our indicator is tailored to identify persistently distressed firms and is, in this regard, aligned with the spirit of the aforementioned laws and paves the way for the implementation of target-specific and evidence-based policies.

Supplementary material for Chapter 4

6 Robustness checks: changing the threshold

Here, we look at what would happen if we would set a different threshold for the failure probability in order for a firm to be identified as a *zombie*. In particular, combining the 8-th and 9-th decile of probability of failure for each firm, for three consecutive years, we were able to identify five different probability levels (i.e., *high, medium-high, medium, medium-low, low*).

Table 27: *Zombies* in 2011

Predicted Risk	Number of Firms	Percentage	Sunk Added Value	Sunk Employment
High Risk	7,278	3.85%	89,853	4.40
Medium-High Risk	4,600	2.44%	232,508	6.91
Medium Risk	3,567	1.89%	254,970	7.32
Medium-Low Risk	3,435	1.82%	415,929	9.76
Low Risk	2,721	1.44%	392,771	9.74
Others	167,195	88.56%	1,130,417	16.91

Note: sunk added value and sunk employment are referred to the average zombie's added value (in Euro) and employment for those firms that failed in the subsequent years. The results are compared with the ones of firms that are not classified as zombies.

These hazard levels rank from high-risk firms (namely, active firms with failure probability above the 9-th decile for three consecutive years), to low-risk firms (namely, active firms with failure probability between the 8-th and 9-th decile for three consecutive years). Such hazard levels catch the risk for a certain firm to be a *zombie* based on its survival probabilities and its failure to exit the market.

Figure 42: *Zombie firms* (8th decile) and GDP

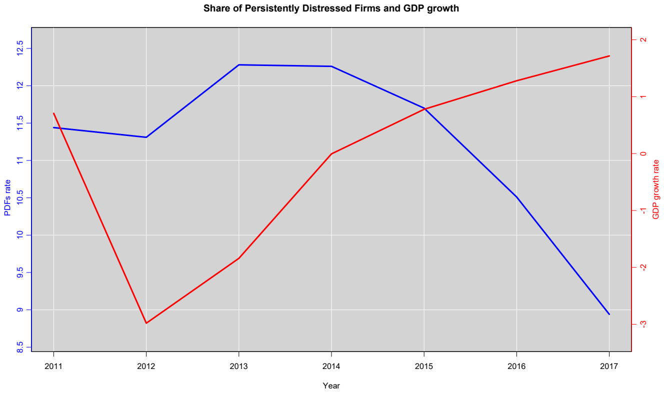


Figure 42 shows the share of *zombies* and the GDP growth rate. Table 28 depicts the selected predictors from LOGIT-LASSO and Table 29 shows the coefficient of the post-logistic regressions for the selected predictors in the time span 2011-2017.

Table 28: Ranking of predictors by year (*zombies* 8th decile)

Rank	2017	2016	2015	2014	2013	2012	2011
1	Control*	Control*	Control*	Control*	Control*	Control*	Control*
2	Size-Age*	Size-Age*	Size-Age*	FCI*	FCI*	FCI*	FCI*
3	ICR*	ICR*	FCI*	Size-Age*	Size-Age*	Size-Age*	Size-Age*
4	FCI*	FCI*	LR*	ICR	LR*	LR*	Solvency*
5	Negative AV*	Negative AV*	ICR*	Negative AV*	ICR*	Solvency*	LR*
6	Solvency*	BID*	Negative AV*	BID*	Innovation*	Negative AV *	Trademark*
7	BID*	LR*	BID*	LR*	Misallocation*	Trademark*	ICR*
8	LR*	Profitability*	LR*	Innovation*	TFP*	Innovation*	Negative AV*
9	CI	CI	Profitability*	Trademark*	BID*	TFP*	Innovation
10	TFP	Solvency*	CI	Profitability*	Negative AV*	BID*	Net Income

Note: the ranking refers to the first ten ranked financial predictors by logit-lasso. The ascetics indicates those variables that are selected from the rigorous logit-lasso as well.

Table 29: Post-logistic regressions with predictors selected with rigorous logit-lasso (*zombies* 8th decile)

	(2017)	(2016)	(2015)	(2014)	(2013)	(2012)	(2011)
Control	-5.503***	-5.870***	-5.551***	-5.293***	-4.872***	-4.212***	-4.072***
	(-51.65)	(-47.75)	(-46.69)	(-42.95)	(-29.45)	(-43.27)	(-31.30)
Size-Age	0.639***	0.884***	1.075***	1.028***	0.857***	0.541***	0.714***
	(11.63)	(15.29)	(18.95)	(17.52)	(20.03)	(19.15)	(18.74)
FCI	0.108	0.702***	1.303***	1.151***	1.570***	1.576***	2.152***
	(1.54)	(8.67)	(12.61)	(13.01)	(9.37)	(24.68)	(15.40)
ICR	1.013***	1.005***	0.560***	0.982***	0.705***	0.301**	0.439***
	(21.98)	(17.85)	(7.17)	(20.09)	(7.36)	(2.68)	(7.18)
Solvency	-0.00770***	-0.00309***		0.00156	-0.00302*	-0.0120***	-0.0164***
	(-9.38)	(-3.79)		(1.65)	(-2.27)	(-14.96)	(-14.96)
BID	0.455***	0.566***	0.410***		0.432***	0.396***	
	(17.87)	(19.37)	(14.27)		(14.24)	(8.65)	
Negative AV	0.604***	1.133***	1.236***	0.667**	0.482	1.544***	0.850***
	(5.21)	(13.31)	(13.32)	(2.95)	(1.51)	(10.55)	(5.86)
Profitability		0.282***	0.456***	0.436***		0.219*	
		(4.46)	(5.86)	(3.94)		(2.25)	
Net income		-3.57e-08***	-2.39e-08*				
		(-4.04)	(-2.56)				
LR	0.0463***	0.128	-1.033**	-6.001***	-4.951***	-2.010***	-1.865***
	(6.11)	(0.53)	(-2.70)	(-12.67)	(-10.62)	(-7.04)	(-6.74)
Innovation			-0.562***	-0.701***	-0.869***	-0.834***	
			(-9.18)	(-6.61)	(-8.32)	(-7.80)	
TFP				-0.00916*	-0.0212***	-0.0227***	
				(-2.38)	(-4.96)	(-5.64)	
Trademark				-0.648***	-0.454***	-1.086***	-1.509***
				(-7.34)	(-3.94)	(-10.91)	(-25.98)
Misallocation					0.314**	0.185**	
					(2.98)	(3.01)	
Sustainability					3.830*		
					(2.13)		
Constant	5.820***	8.161***	10.29***	10.34***	8.430***	4.882***	6.770***
	(10.35)	(13.00)	(16.92)	(16.63)	(17.08)	(14.92)	(17.56)
Year FE	(YES)	(YES)	(YES)	(YES)	(YES)	(YES)	(YES)
Sector FE	(YES)	(YES)	(YES)	(YES)	(YES)	(YES)	(YES)
Observations	78635	80365	82459	78549	44191	51733	76273

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Conclusions

Machine learning algorithms have advanced to become effective and prominent tools across a broad-spectrum of scientific fields and applications. The reasons behind the surge in the usage of machine learning are manifold and are connected to the major technological innovations introduced in the last decades. Advances in data storage, data transfer and data processing, and the availability of large data sources to be analyzed called for novel, more powerful, computational tools. This staggering growth in the size of datasets was mainly driven by the increase in the amount of data produced through novel technologies.

The early adopters of machine learning techniques were mainly computer scientists (mostly, on the application side) and statisticians/applied mathematicians (mostly, on the theoretical side). However, as these technologies have become more widespread, their influence has gone far beyond the boundaries of these fields. Indeed, the application of machine learning in social sciences has given rise to a completely novel field of investigation named "computational social science".

Machine learning techniques can capture hidden patterns in the data by directly learning from them, without the usual restrictions of model-based statistical methods, due to their model-free, data driven nature (Breiman et al., 2001). These features are particularly central as researchers are not bounded to theory-based methodologies and assumptions, but they can "let the data speak" and, in some cases, let the data drive their theories. This does not mean that theories and assumptions are undermined in the novel “machine learning era” (Sejnowski, 2018). Quite the contrary, machine learning alone cannot provide an insight on the underlying assumption of causal inference settings (Dominici & Mealli, 2020). Also, a theory-driven choice regarding which machine learning technique best fits the application at hand can lead to a great increase in the predictive performance of the model.

The main aim of this Dissertation is to extend the causal and predictive machine learning frameworks through an application-oriented perspective. Indeed, the methods developed provide answers to relevant empirical questions. In the first two Chapters, we bridge several gaps in the field of interpretable causal machine learning for heterogeneous effects estimation and discovery in observational studies. In the last two Chapters, we adapt the usual predictive machine learning framework to account for robust and theory driven predictions. Table 30 provides an overview of the main empirical contributions of the Dissertation.

1 Methodological contribution

In the first Chapter, we introduced a novel, Bayesian machine learning methodology to discover and estimate heterogeneous effects in quasi-experimental settings. Analysis of heterogeneous treatment effects are usually conducted for subgroups defined a-priori (Cook et al., 2004). A-priori choice of the subgroups of analysis is required[14] to avoid possible cherry-picking problems: i.e., researchers reporting the causal results only for subgroups with extremely high/low treatment effects. However, defining subgroups a priori has some drawbacks: it requires a good understanding of the treatment effect (from previous literature) and, it could miss unexpected subgroups. Moreover, few methodologies have been proposed to deal with heterogeneous effect discoveries in observational studies in the presence of a broken randomization scenario (Bargagli-Stoffi & Gnecco, 2020). In this Chapter, we account for these shortcomings, by proposing a data-driven methodology to discover and

[14] This is, for instance, the case of the protocols issued by the US Food and Drug Administration.

estimate heterogeneous effects in observational studies with imperfect compliance. Such a methodology, that we name Bayesian Causal Forest with Instrumental Variable (BCF-IV), outperforms seminal methodologies proposed in the field (Athey et al., 2019; Bargagli-Stoffi & Gnecco, 2018) both in discovering the effects and in precisely estimating them in a simulated scenario. On top of that, BCF-IV has desirable statistical properties such as consistency and asymptotic normality of the conditional estimators.

The second Chapter deals with a widely known issue of machine learning techniques: their potential lack of interpretability. As causal machine learning methodologies are providing more and more accurate predictions of individual treatment effects, at the same time their complexity is increasing as well, leading to a complete lack of interpretability in some cases (Doshi-Velez & Kim, 2017). We argue that heterogeneous effects are central for targeted policies, but in high-stakes scenarios simple, yet consistent, causal rules should be provided to policy-makers and stakeholders. Indeed, interpretable causal machine learning can promote a deeper understanding about how the subpopulations to be targeted were discovered. This, in turn, provides a higher trust in these methodologies both with regard to policy-makers enacting the policies, and to constituent subject to their implementation. In this spirit, we propose an algorithm that aims at a higher level of "accountability" in the discovery and estimation of causal effects. Beside leading to a higher level of interpretability, the proposed method accounts for several limitations of the most interpretable causal machine learning methodologies introduced to this day. For instance, tree based methods provide a high level of interpretability but suffer of three main limitations: (i) they are unstable; (ii) they are able to discover just a limited number of heterogeneous subpopulations; (iii) they often fail do distinguish between effect modifiers and confounders. The proposed Causal Rule Ensemble accommodates for these limitations by providing a stable algorithm for the discovery of a large set of rules controlling for confounding factors. In a series of simulations, we showed how the proposed method outperforms state-of-the-art algorithms while providing consistent results.

While in the first two Chapters we deal with issues related to causality in general, and heterogeneous effects, in the last two we address policy prediction problems (Kleinberg et al., 2015).

In the third Chapter, we introduce a novel sensitivity of predictions analysis. We argue that one potential problem arising from the extensive usage of machine learning to address inherently predictive problems, could be an underlying instability in the predictive model. This issue is central as high-stakes predictive models should not just display high performance, but also stability. This is utterly true in applications where there is no real performance benchmark, and researchers cannot rely on previous contributions for a comparative analysis. In our contribution, we define a predictive model to be stable if its predictions and its performance are not widely changed by the inclusion of a new synthetic variable with a high explanatory power in the set of predictors. We name this simulation-based analysis *sensitivity of predictions*. Sensitivity of predictions can inform practitioners and policy-makers on the quality of their predictive model. In particular, the main questions that it seeks to answer are the following: *"is the machine learning model getting enough valuable signal from the data?"* and *"how much would an unobserved predictor impact the model's prediction and its performance?"*. Such questions are firstly investigated in this Chapter.

The contribution of the fourth Chapter is not methodological in the strictest sense. Indeed, it relates to a wider definition of *methods* as procedures for accomplishing a predictive task. We start by asking ourselves if there is a way to improve the predictive power of state-of-the-art techniques by incorporating theoretical intuitions in the predictive model for the phenomenona being investigated. Indeed, the application we deal with has been addressed many times in the literature on machine learning for firm dynamics, and many different methods have been proposed (Bargagli-Stoffi, Niederreiter, et al., 2020). We find that a procedure that starts from the economic problem at hand and, consequently, chooses the model that could best fit the specific problem (from a theoretical perspective), provides a great improvement in the predictive performance. Of course, such findings are specific to the context of our application. However, we argue that the proposed procedure shows that it is not always the case that the most complex method provides the best performance, while it can often be the case that an application oriented choice can lead to the best results.

2 Policy implications

The empirical research question that we investigate in Chapter 1 concerns the estimation of the heterogeneous effects of additional school funding on students' performance:

> **RQ1:** ***What kinds of students most benefit from additional school funding?***

Such question is relevant from an empirical perspective as it relates, on one side, with the effectiveness of educational funding policies, and on the other, with student's differential human capital development. The proposed BCF-IV methodology enables us to investigate the effects of additional school funding promoted by the Flemish Ministry of education starting from 2002. We study the performance of 135,682 treated and untreated students, accounting for the universe of pupils in Flanders in the school year 2010/2011. We find that the overall effects are slightly positive but not significant. This finding is in line with the most recent literature on students' performance both in Flanders (De Witte et al., 2018) and in a wider cross-country perspective (Hanushek et al., 2016; Hanushek & Woessmann, 2017). However, we find evidence of strong effect variation. Indeed, students in schools with less senior/younger principals and teachers show higher, positive, significant effects of the funding on their performance. This insight is particularly important, not just because it is the first time that heterogeneous effects are investigated in the context of funding effectiveness. In fact, the underlying mechanism that we discover informs policy-makers in multiple ways. Firstly, it highlights the need for additional funds to schools with younger teachers and principals. In many cases, senior teachers and principals select themselves out of the most disadvantaged schools[15], creating more vacancies in such schools. Therefore, younger teachers and principals lack a real choice but to start working in the most disadvantaged schools. Hence, supporting younger teachers and principals is central as it leads to higher returns in terms of students' performance, and it could help bridging the performance gap between advantaged and disadvantaged schools. Secondly, the discovered heterogeneity in the effect could hint at differential motivation levels between younger and more senior teachers and principals, that could, in turn, have an impact on their performance. Ololube (2006) favours this hypothesis, as he finds the motivation enhances teachers' productivity, with potentially beneficial effects on students' performance.

The research question that we investigate in the application section of Chapter 2 is:

> **RQ2:** ***What characteristics make people most vulnerable to the negative effects of higher levels of pollution?***

We use data on Medicare beneficiaries in New England between 2000 and 2006 and pollution data on the average two-year exposure (2000-2001) to $PM_{2.5}$ in the same regions. The outcome of our investigation is the five-year mortality rate measured between 2002 and 2006, while the characteristics of the individuals that we consider are unit level covariates - sex (male,female), age (65-70, 71-75, 76-80, 81-85, 86+), race (white, non-white), and Medicaid eligibility (eligible, non-eligible). Medicaid eligibility is considered as a proxy variable indicating low socio-economic status. Moreover, we introduce 9 ZIP code-level or 2 county-level covariates (average temperature, humidity, body-mass-index, smoker rate, Hispanic population rate, median household income, median value of housing, % of people below the poverty level, % of people below high school education, % of owner occupied housing and population density). These variables are used both to control for potential confounding bias as well as effect variation drivers. We find that, controlling for

[15] We refer to Chapter 1 for a the detailed definition of what "disadvantaged" and "advantaged" mean in this context.

confounding effects, Medicaid eligible white people aged between 81-85 and white people above 80 are most at risk. These results are consistent with the findings of K. Lee et al. (2018) and are validated through a sensitivity analysis. Obviously, it is critical to understand which people are more vulnerable to the effect of pollution. Indeed, this analysis can inform policy-makers on who should be targeted and what policies can be designed to prevent or mitigate the negative effects of pollution on health. Moreover, our findings are informative on the areas where the efforts to reduce pollution should be greater. The higher the density of older, low income people, the more restrictive the environmental policies should be.

Chapter 3 investigates whether or not a predictive model for student's financial literacy scores with input data on student-level (age, socio-economic status, achievements, attitude, among the others) and school-level characteristics (school size, classes, size, teachers' professional development, among the others) provides stable and robust predictions:

> **RQ3:** ***How sensitive are the predictions of student financial literacy scores to a potentially unobserved predictor with high explanatory power?***

We use the 2015 data of the OECD's Program for International Student Assessment (PISA) for Belgium to train our model. Through the novel *sensitivity of prediction analysis*, we find that the model's predictions are fairly invariant to the inclusion of a new synthetic predictor with a high explanatory power. This finding informs us that the predictors provide enough signal to the predictive model. On top of that, it informs policy-makers on the viability of the model that reaches an adjusted R^2 out-of-sample performance of 73%. The model is used to predict the financial literacy scores for a region of Belgium (Wallonia) where these scores are not observed, using the same set of input variables. This model could be used to target students with low predicted financial literacy scores with additional teaching material to boost their financial knowledge. Such intervention would be highly valuable as the literature demonstrates the centrality of financial literacy to an active participation in society (OECD, 2017b). Our results show that the probability of having low financial literacy score is largest for students with lower scores on reading and math. This evidence is corroborated by Mancebón et al. (2019) and highlights how the development of financial abilities could be mediated by mathematical skills. Moreover, we find that students with the lower predicted financial literacy scores are often students from families where the school language is not spoken at home. Also, the educational background of parents is another important predictor. These observations are in line with existing literature suggesting that more vulnerable groups in terms of low financial literacy scores are often situated among those with a lower socioeconomic background (De Beckker et al., 2019; Gramaţki, 2017; Riitsalu & Põder, 2016). Even if these findings do not indicate a causal relationship between these characteristics and financial literacy scores, they provide hints to policy-makers on possible lines of action to boost students' financial literacy. Indeed, we argue that policy-makers should opt a 360° approach to boost students' financial literacy scores. First, they could use our machine learning methodology to spot students with lower predicted scores. Then, for such students they could implement a targeted learning approach not only focused at improving financial literacy, but also math and reading. Additionally, our methodology could be used to target student at a very early stage, in a way that could boost the positive results and externalities of the interventions.

While in the previous Chapter, we dealt with a novel setting where machine learning techniques were either employed for the first time (Chapter 1 and 3) or,very rarely (Chapter 2), in Chapter 4 we deal with an application that was extensively investigated in the literature. As shown by Bargagli-Stoffi, Niederreiter, et al. (2020), there are various applications of machine learning predictive power in the field of firm dynamics. Hence, the goal of the fourth Chapter is to understand if and how we can possibly outperform previously proposed methodologies by tackling the predictive problem from a different perspective:

RQ4: ***Can economic intuition boost the machine learning predictions of firms' distress and provide insight on non-***

viable firms that do, however, still survive?

The contribution of our Chapter is relevant from a policy perspective in at least three ways. First, we show that accounting for failure to release financial information can improve credit risk prediction in a staggering way. Second, we propose a novel machine learning-based definition of *zombie firms*. This definition can be used to identify firms that are in a persistently distressed position. Third, the method that we implement allows us to validate the current, deterministic indicators used to assess firm's credit risk and financial distress. We find that *financial constraint indicators*, like the ones proposed by Nickell and Nicolitsas (1999) and Hadlock and Pierce (2010), and *zombie* indicators such as *interest coverage ratio* (Bank of England, 2013), *negative added value* (Bank of Korea, 2013), *profitability* (Schivardi et al., 2017), liquidity return and solvency show the best predictive performance. This is particularly relevant as the European Union (under the Directive 2012/30/EU), and other European countries, like Italy in particular (see Law October 19th 2017, n. 155), are re-calibrating in their business failure legislation to include in the financial assessment of firms' viability indicators tailored to signal the crises of firms in an early stage. In this spirit, our indicator is tailored to identify persistently distressed firms and is aligned with the aforementioned laws, paving the way for the implementation of target-specific and evidence-based policies.

To conclude, machine learning provides an incredibly useful set of tools to improve the effectiveness of public policies in many ways. As shown in this Dissertation, on the one side, causal machine learning could be used to either design optimal policies, or to boost the effect of existing policies. On the other side, predictive machine learning could be employed to gather preemptive insights on potential groups to be targeted. We argue that policy-makers should increasingly adopt an evidence-bases approach. In this spirit, they should set up teams of statisticians, health, social and data scientists with the aim of gathering - through the application of novel methodologies in the fields of causality and machine learning - real-world evidence on the actual, or potential, effects of public interventions.Furthermore, with a view to make any data analysis possible, engineers and computer scientists should be hired to improve data collection, transfer and storage. The costs related to these teams would be outweighed by the benefits in the effectiveness of policies, or by the decrease in their related costs. On top of that, policymakers could use these novel technologies to design personalized interventions and to reach higher levels of accountability.

3 Outlines for further research

In the first Chapter, we presented evidence of strong teacher and principal related drivers of funding effectiveness on students' performance. However, the investigation of the true causal channels is beyond the goals of this Chapter and are left to further investigation where more granular teachers' and principals' characteristics are available. We argue that providing such an insight on the drivers of funding effectiveness is not only important to boost students' performance, but also important to identify which are the underlying features of students, teachers, principals and schools that lead to an optimal policy allocation. Furthermore, the method that we develop, BCF-IV, could be extended, connecting it with the *Generalized Random Forest* algorithm (Athey et al., 2019). Such an ensemble algorithm could be created to obtain a novel technique that combines the small and large sample properties of both the methodologies. Indeed, as shown by Van der Laan et al. (2007) it is often the case that the combination of multiple machine learning techniques results in an ensemble method that usually outperforms the single methodologies.

In Chapter 2, we saw that older, white people with low income are more vulnerable to the exposure to higher levels of pollution. Such evidence can provide a simple rule to implement policies targeted at preventing or mitigating the negative effects of pollution for this demographic. However, the data that we used do not enable us to understand which are the more granular drivers of this higher hazard. A very interesting line for further research would be to investigate, with an ad-hoc study, whether the drivers of vulnerability can be found in a higher morbidity, and what are the diseases that are most associated with higher increases in the mortality rates. On the application side, it would be promising to investigate the effects of environmental policies: e.g., does a decrease in pollution levels affect in different ways different subgroups of people? On top of that, a limitation of our study is that we consider pollution exposure as a binary variable, dichotomizing the levels using an exogenously set threshold. We argue that a very promising methodological contribution would be to extend our setting from a binary to a continuous treatment scenario.

The third- and fourth- Chapter dealt with applications of machine learning in the field of policy prediction problems. Hence, a straightforward extension of these works would be to run a causal inference analysis regarding the effects of the predictors on the outcome. Such analyses are not straightforward to implement because they need a proper experimental or quasi-experimental scenario to draw proper causal inference. However, we argue that a very interesting extension of our work would go in the direction of assessing how exogenous increases in reading and math scores would affect financial literacy. Indeed, programs targeted at enhancing students' financial literacy could be ineffective if they do not provide to students with the tools to improve their knowledge in math and their reading abilities. This could hold true as we find that reading and math scores are good predictors of financial literacy. Moreover, it would be of interest to use a methodology similar to the one that we proposed to shift from spatial predictions (i.e., predicting scores in regions where they are not observed) to temporal predictions (i.e., predicting scores one, two, three years from now based on the data that we observe today).

Regarding possible further research related to the fourth Chapter, we believe that there is the need for a causal analysis on how a revision of national regulations for bankruptcy laws may help reduce the presence of zombie firms. Moreover, we argue that researchers should further investigate the phenomenon of *zombie lending* (i.e., banks loaning money to non-viable, financially distressed enterprises). Indeed, a better allocation of credit could avoid crowding out of financial resources at the expenses of more viable firms with profitable projects of investments. To the best of our knowledge, scant attention has been devoted to the public sector and the criminal sector as drivers of the *zombie* phenomenon. Such investigations would be of great importance to reduce possible public misallocation of credit, and to furnish to public authorities an insight on how important the criminal component of *zombie* lending is. We leave this to further research.

Thus far, we highlighted further lines of research directly connected with the works presented in each Chapter of this Dissertation. However, we want to take a small step away from our works and look, from a bigger perspective, at few of the most promising directions of future research in the field of applied machine learning. From a *predictions-in-policies* perspective, we argue that in the next few years more effort should be made in providing higher levels of interpretability. A big challenge will be to *translate* very complex machine learning tools (e.g., deep artificial neural networks) into more interpretable, and accountable algorithms. While predictive machine learning has been used almost in any possible context - from algorithms that spot people's feelings to Netflix movies' suggestions - we argue that there are many more *landscapes* to be explored in the field of causal machine learning. Promising lines of research in this field regard extending the discovery and estimation of causal effects to scenarios where the treatment is continuous (i.e., exposure-response curves), or where there are interaction patterns between the units that would introduce strong biases in the causal estimands (i.e., social networks). However, the real "shot-to-the-moon" in both fields is to investigate in depth the theoretical properties and empirical performances of each method to provide a better guidance to practitioners regarding which techniques to use in any specific context. With this Herculean effort, the application of machine learning will move from an *art* (Peng & Matsui, 2015) to higher standards of reproducibility.

In conclusion, it requires a deep understanding of the research question under investigation to design, develop or adjust a machine learning methodology in social and health sciences scenarios. This holds true also when it comes to understanding in depth the insights provided by novel machine learning techniques. We are far from machines overturning the importance of scholars' abilities such as properly designing a causal evaluation or, deeply understanding the phenomenon being studied. The most recent developments of applied machine learning literature, as well as this Dissertation's contributions, provide insights on the fact that the most effective usage of novel technologies can be reached when *intelligent machines are driven by expert human beings*.

Table 30: The dissertation's empirical contributions

	Chapter	Research Question	Findings	Policy implications	Further Research
Part 1	Chapter 1	What kinds of students most benefit from additional school funding?	Students in schools with less senior/younger principals and teachers show positive, significant effects	Increase funding to schools that benefit the most from additional financial support	Investigate the causal mechanism that drives effects variation at a more granular level
	Chapter 2	What characteristics make people most vulnerable to the negative effects of higher levels of pollution?	Medicaid eligible white people aged between 81-85 and white people above 80 are most at risk	Targeted policies to prevent or mitigate the negative effects of pollution for the most vulnerable subgroups	Investigate possible health related drivers of the heterogeneous effects (e.g. morbidity)
Part 2	Chapter 3	How sensitive are the predictions of student financial literacy scores to a potentially unobserved predictor with high explanatory power?	Predictions are fairly invariant to the inclusion of a new synthetic predictor	Provide students with low predicted financial literacy scores with additional teaching material at an early stage	Shift from inter-spacial to inter-temporal predictions of financial literacy scores
	Chapter 4	Can economic intuition boost the machine learning predictions of firms' distress and provide insight on non-viable firms that do, however, still survive?	Accounting for failure to release financial information can improve credit risk predictions and the discovery of *zombie firms*	Prevent banks from indulging in zombie lending, and revise national regulations for bankruptcy laws	Causal investigation of possible public and criminal drivers of persistently distressed firms' survival

Note: the policy implications illustrated in the Table are just some of the various, possible implications connected to our empirical findings. The same holds true for further lines of research. Please refer to the related Chapters for more detailed discussions.

Bibliography

Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, *83*(6), 2411–2451.

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2019). Lassopack: Model selection and prediction with regularized regression in stata. *arXiv preprint arXiv:1901.05397*.

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, *94*, 164–184.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589–609.

Altman, E. I. (2000). Predicting financial distress of companies: Revisiting the z-score and zeta models. *Stern School of Business, New York University*, 9–12.

Andrews, D., McGowan, M. A., Millot, V., Et al. (2017). *Confronting the zombies: Policies for productivity revival* (tech. rep.). OECD Publishing.

Angrist, J. D., Imbens, G., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Athey, S. (2019). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda* (pp. 507–547). University of Chicago Press. https://doi.org/https://doi.org/10.7208/chicago/9780226613475.001.0001

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., & Imbens, G. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Athey, S., & Imbens, G. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Athey, S., & Wager, S. (2017). Efficient policy learning. *arXiv preprint, arXiv:1702.02896*.

Bank of England. (2013). *Inflation report* (tech. rep.). Bank of England, London.

Bank of Korea. (2013). *Financial stability report* (tech. rep.). Bank of Korea, Seoul.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417.

Bargagli-Stoffi, F. J., & Gnecco, G. (2020). Causal tree with instrumental variable: An extension of the causal tree framework to irregular assignment mechanisms. *International Journal of Data Science and Analytics*, *9*, 315–337.

Bargagli-Stoffi, F. J., De Beckker, K., De Witte, K., & Maldonado, J. E. (2020). Assessing sensitivity of predictions. a novel toolbox for machine learning with an application on financial literacy. *Working paper*.

Bargagli-Stoffi, F. J., De Witte, K., & Gnecco, G. (2019). Heterogeneous causal effects with imperfect compliance: A novel bayesian machine learning approach. *arXiv preprint arXiv:1905.12707*.

Bargagli-Stoffi, F. J., & Gnecco, G. (2018). Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms. *In Proceedings of the 5th IEEE Conference in Data Science and Advanced Analytics*, 1–10.

Bargagli-Stoffi, F. J., Niederreiter, J., & Riccaboni, M. (2020). Supervised learning for the prediction of firm dynamics (S. Consoli, D. Reforgiato Recupero, & M. Saisana, Eds.). In S. Consoli, D. Reforgiato Recupero, & M. Saisana (Eds.), *Data science for economics and finance: Methodologies and applications*. Springer.

Bargagli-Stoffi, F. J., Riccaboni, M., & Rungi, A. (2020). Machine learning for zombie hunting. firms' failures and financial constraints. *Working paper*.

Bargagli-Stoffi, F. J., Tortu, C., & Forastiere, L. (2020). Heterogeneous treatment and spillover effects under clustered network interference. *Working paper*.

Barnes, R., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy*, *43*(5), 772–777.

Behr, A., & Weinblat, J. (2017). Default patterns in seven eu countries: A random forest approach. *International Journal of the Economics of Business*, *24*(2), 181–222.

Bellman, R. E. (1961). *Adaptive control processes: A guided tour*. Princeton University Press.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Belloni, A., Chernozhukov, V., Hansen, C., & Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, *34*(4), 590–605.

Belloni, A., Chernozhukov, V., & Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, *34*(4), 606–619.

Ben-Michael, E., Feller, A., & Rothstein, J. (2018). The augmented synthetic control method. *arXiv preprint arXiv:1811.04170*.

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning*, *106*(7), 1039–1082.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of political economy*, *81*(3), 637–654.

Bonello, J., Brédart, X., & Vella, V. (2018). Machine learning models for predicting financial distress. *Journal of Research in Economics*, *2*(2), 174–185.

Brédart, X. (2014). Bankruptcy prediction model using neural networks. *Accounting and Finance Research*, *3*(2), 124–128.

Breiman, L. (1984). *Classification and regression trees*. New York, New York, Routledge.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Breiman, L. Et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199–231.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. *Belmont, CA: Wadsworth & Brooks*.

Buuren, S. v., & Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 1–68.

Caballero, R. J., Hoshi, T., & Kashyap, A. K. (2008). Zombie lending and depressed restructuring in japan. *American Economic Review*, *98*(5), 1943–77.

Calligaris, S., Del Gatto, M., Hassan, F., Ottaviano, G. I., & Schivardi, F. (2016). *Italy's productivity conundrum. a study on resource misallocation in italy* (tech. rep.). Directorate General Economic and Financial Affairs (DG ECFIN), European Commission.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015). Rdrobust: An r package for robust nonparametric inference in regression-discontinuity designs. *R Journal*, *7*(1), 38–51.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124–27.

Chandra, D. K., Ravi, V., & Bose, I. (2009). Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with Applications*, *36*(3), 4830–4837.

Chen, J., & Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, *107*(5), 261–65.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Chipman, H. A., George, E. I., McCulloch, R. E., Et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

Cleofas-Sánchez, L., Garcıa, V., Marqués, A., & Sánchez, J. S. (2016). Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing*, *44*, 144–152.

Cockx, B., Lechner, M., & Bollens, J. (2019). Priority to unemployed immigrants? a causal machine learning evaluation of training in belgium. *arXiv preprint arXiv:1912.12864*.

Cohen, P., West, S. G., & Aiken, L. S. (2014). *Applied multiple regression/correlation analysis for the behavioral sciences*. Psychology Press.

Coleman, J. S. (1966). Equality of educational opportunity. *Washington DC: US Government Printing Office*, 1–32.

Cook, D. I., Gebski, V. J., & Keech, A. C. (2004). Subgroup analysis in clinical trials. *Medical Journal of Australia*, *180*(6), 289–291.

Cordero, J. M., & Pedraja, F. (2018). The effect of financial education training on the financial literacy of Spanish students in PISA. *Applied Economics*, *51*(16), 1679–1693. https://doi.org/10.1080/00036846.2018.1528336

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, *90*(3), 389–405.

Danenas, P., & Garsva, G. (2015). Selection of support vector machines based classifiers for credit risk domain. *Expert Systems with Applications*, *42*(6), 3194–3204.

Davis, J., & Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, *107*(5), 546–50.

De Beckker, K., De Witte, K., & Van Campenhout, G. (2019). Identifying financially illiterate groups: An international comparison. *International Journal of Consumer Studies*, *43*(5), 490–501.

De Witte, K., Smet, M., & D'Inverno, G. (2018). The effect of additional resources for schools with disadvantaged students: Evidence from a conditional efficiency model. *Steunpunt Sono Research Report*.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506.

Devijver, P. A., & Kittler, J. (1982). *Pattern recognition: A statistical approach*. Prentice hall.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., & Schwartz, J. (2016). Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environmental science & technology*, *50*(9), 4712–4721.

Ding, P., Feller, A., & Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*, *114*(525), 304–317.

Dominici, F., & Mealli, F. (2020). From controlled to undisciplined data: Estimating causal effects in the era of data science using a potential outcome framework. *Harvard Data Science Review, to appear*.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dubé, J.-P., & Misra, S. (2017). *Scalable price targeting* (tech. rep.). National Bureau of Economic Research.

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from kenyan primary schools. *Journal of public Economics*, *123*, 92–110.

Eberts, R. W., & Stone, J. A. (1988). Student achievement in public schools: Do principals make a difference? *Economics of Education Review*, *7*(3), 291–299.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans* (Vol. 38). Philadelphia, PA: Society for Industrial; Applied Mathematics.

Fantazzini, D., & Figini, S. (2009). Random survival forests models for sme credit risk measurement. *Methodology and computing in applied probability*, *11*(1), 29–45.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

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*(177), 1–81.

Foster, J. C., Taylor, J. M., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30*(24), 2867–2880.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(9), 1189–1232.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916–954.

Gal, P. N. (2013). Measuring total factor productivity at the firm level using oecd-orbis.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). Bayesian data analysis (vol. 2). *Boca Raton, FL: Chapman*.

Gentilucci, J. L., & Muto, C. C. (2007). Principals' influence on academic achievement: The student perspective. *NASSP bulletin*, *91*(3), 219–236.

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review*, *100*(2), 250–55.

Gopinath, G., Kalemli-Özcan, Ş., Karabarbounis, L., & Villegas-Sanchez, C. (2017). Capital allocation and productivity in south europe. *The Quarterly Journal of Economics*, *132*(4), 1915–1967.

Gottard, A., Vannucci, G., & Marchetti, G. M. (2020). A note on the interpretation of tree-based regression models. *Biometrical Journal*.

Gramaţki, I. (2017). A comparison of financial literacy between native and immigrant school students. *Education Economics*, *25*(3), 304–322. https://doi.org/10.1080/09645292.2016.1266301

Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opinion Quarterly*, *76*(3), 491–511.

Grimmer, J., Messing, S., & Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, *25*(4), 413–434.

Gruszczyński, M. (2019). On unbalanced sampling in bankruptcy prediction. *International Journal of Financial Studies*, *7*(2), 28.

Hadlock, C. J., & Pierce, J. R. (2010). New evidence on measuring financial constraints: Moving beyond the kz index. *The Review of Financial Studies*, *23*(5), 1909–1940.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209.

Hahn, P. R., Carvalho, C. M., Puelz, D., He, J., Et al. (2018). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, *13*(1), 163–182.

Hahn, P. R., Dorie, V., & Murray, J. S. (2018). *Atlantic causal inference conference (acic) data analysis challenge 2017*.

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, *143*(1), 29–36.

Hansen, C., Hansen, C., Rastin, M., & Pia, M. (2018). Predicting distresses using deep learning of text segments in annual reports. *Danmarks Nationalbank Working Paper*, (130).

Hansen, C., & Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics*, *182*(2), 290–308.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *The Economic Journal*, *113*(485), F64–F98.

Hanushek, E. A., Machin, S. J., & Woessmann, L. (2016). *Handbook of the economics of education*. Elsevier.

Hanushek, E. A., & Woessmann, L. (2017). School resources and student achievement: A review of cross-country economic research, In *Cognitive abilities and educational outcomes*. Springer.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of public economics*, *95*(7-8), 798–812.

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2016). Counterfactual prediction with deep instrumental variables networks. *arXiv preprint, arXiv:1612.09596*.

Hassoun, M. H. Et al. (1995). *Fundamentals of artificial neural networks*. MIT press.

Heo, J., & Yang, J. Y. (2014). Adaboost based bankruptcy forecasting of korean construction companies. *Applied soft computing*, *24*, 494–499.

Hernández, B., Raftery, A. E., Pennington, S. R., & Parnell, A. C. (2018). Bayesian additive regression trees using bayesian model averaging. *Statistics and Computing*, *28*(4), 869–890.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Holmlund, H., & Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher? *Labour Economics*, *15*(1), 37–53.

Hosaka, T. (2019). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert systems with Applications*, *117*, 287–299.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

Ichino, A., Mealli, F., & Nannicini, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of applied econometrics*, *23*(3), 305–327.

Imbens, G., & Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, *64*(4), 555–574.

Imbens, G., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jackson, C. K. (2018). *Does school spending matter? the new literature on an old question* (tech. rep.). National Bureau of Economic Research.

Jackson, C. K., Johnson, R. C., & Persico, C. (2015). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *The Quarterly Journal of Economics*, *131*(1), 157–218.

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2018). Teacher applicant hiring and teacher performance: Evidence from dc public schools. *Journal of Public Economics*, *166*, 81–97.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Tree-based methods, In *An introduction to statistical learning*. Springer.

Johnson, M., Cao, J., & Kang, H. (2019). Detecting heterogeneous treatment effect with instrumental variables. *arXiv preprint arXiv:1908.03652*.

Kalemli-Ozcan, S., Sorensen, B., Villegas-Sanchez, C., Volosovych, V., & Yesiltas, S. (2015). *How to construct nationally representative firm level data from the orbis global database* (tech. rep.). National Bureau of Economic Research.

Kallus, N. (2018). Balanced policy evaluation and learning, In *Advances in neural information processing systems*.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (tech. rep.). National Bureau of Economic Research.

Kapelner, A., & Bleich, J. (2015). Prediction with missing data via bayesian additive regression trees. *Canadian Journal of Statistics*, *43*(2), 224–239.

Kapelner, A., & Bleich, J. (2016). Bartmachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software, Articles*, *70*(4), 1–40. https://doi.org/10.18637/jss.v070.i04

Kennedy, A. B. W., & Sankey, H. R. (1898). *The thermal efficiency of steam engines* (Minutes No. 134). Proceedings of the Institution of Civil Engineers.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability, In *Advances in neural information processing systems*.

Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and adaboosted decision tree models. *Economic Modelling*, *36*, 354–362.

Kitagawa, T., & Tetenov, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica*, *86*(2), 591–616.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–95.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*, *37*(4), 331–344.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students'performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, *18*(5), 411–426.

Kotthoff, L. (2016). Algorithm selection for combinatorial search problems: A survey, In *Data mining and constraint programming*. Springer.

Lechner, M. (2019). Modified causal forests for estimating heterogeneous causal effects. *CEPR Discussion Paper No. DP13430*.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Lee, I., & Shin, Y. J. (2019). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*.

Lee, K. C., Han, I., & Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. *Decision Support Systems*, *18*(1), 63–72.

Lee, K., Bargagli-Stoffi, F. J., & Dominici, F. (2020). Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. *Working paper*.

Lee, K., Small, D. S., & Dominici, F. (2018). Discovering effect modification and randomization inference in air pollution studies. *arXiv preprint, arXiv:1802.06710*.

Lee, M.-j. (2009). Non-parametric tests for distributional treatment effect for randomly censored responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(1), 243–264.

Lewis, J. B., & Linzer, D. A. (2005). Estimating regression models in which the dependent variable is based on estimates. *Political Analysis*, *13*(4), 345–364.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766.

Li, F., Mattei, A., & Mealli, F. (2015). Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, 1906–1931.

Liang, D., Lu, C.-C., Tsai, C.-F., & Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, *252*(2), 561–572.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, *113*(522), 626–636.

Linero, A. R., & Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(5), 1087–1110.

Linn, M., & Weagley, D. (2019). Estimating financial constraints with machine learning. *Available at SSRN 3375048*.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Logan, B. R., Sparapani, R., McCulloch, R. E., & Laud, P. W. (2019). Decision making and uncertainty quantification for individualized treatments using bayesian additive regression trees. *Statistical Methods in Medical Research*, *28*(4), 1079–1093.

Loh, W.-Y., Cao, L., & Zhou, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *Wiley In-*

terdisciplinary Reviews: Data Mining and Knowledge Discovery, *9*(5), e1326.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217–224.

Longobardi, S., Pagliuca, M. M., & Regoli, A. (2018). Can problem-solving attitudes explain the gender gap in financial literacy? Evidence from Italian students' data. *Qual Quant*, *52*, 1677–1705. https://doi.org/10.1007/s11135-017-0545-0

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). *Forecasting methods and applications*. John wiley & sons.

Mancebón, M.-J., Ximénez-de-Embún, D. P., Mediavilla, M., & Gómez-Sancho, J.-M. (2019). Factors that influence the financial literacy of young Spanish consumers. *International Journal of Consumer Studies*, *43*, 227–235. https://doi.org/10.1111/ijcs.12502

Masci, C., Paganoni, A. M., & Ieva, F. (2019). Semiparametric mixed effects models for unsupervised classification of italian schools. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(4), 1313–1342.

Mattei, A., & Mealli, F. (2016). Regression discontinuity designs as local randomized experiments. *Observational Studies*, *66*, 156–173.

Mazrekay, D., Schiltz, F., & Titl, V. (2018). Identifying politically connected firms: A machine learning approach.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*(2), 698–714.

McGowan, M. A., Andrews, D., & Millot, V. (2018). The walking dead? zombie firms and productivity performance in oecd countries. *Economic Policy*, *33*(96), 685–736.

Mealli, F., & Rampichini, C. (2012). Evaluating the effects of university grants by using regression discontinuity designs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(3), 775–798.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*(3), 1436–1462.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, *29*(2), 449–470.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology*, *43*(4), 907–912.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Mitchell, T. M. (1997). Machine learning. *Burr Ridge, IL: McGraw Hill*, *45*(37), 870–877.

Moscatelli, M., Narizzano, S., Parlapiano, F., Viggiano, G., Et al. (2019). *Corporate default forecasting with machine learning* (tech. rep.). Bank of Italy, Economic Research and International Relations Area.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Murray, J. S. (2017). Log-linear bayesian additive regression trees for categorical and count responses. *arXiv preprint, arXiv:1701.01503*.

Nalenz, M., & Villani, M. (2018). Tree ensembles with rule structured horseshoe regularization. *The Annals of Applied Statistics*, *12*(4), 2379–2408.

Nethery, R. C., Mealli, F., Dominici, F., Et al. (2019). Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *The Annals of Applied Statistics*, *13*(2), 1242–1267.

Neyman, J. (1923, 1990). On the application of probability theory to agricultural experiments, section 9 [reprinted in Statistical Science, 1990, 5, 463–485]. *Roczniki Nauk Roiniczych*, *10*, 1–51.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, In *Advances in neural information processing systems*.

Nickell, S., & Nicolitsas, D. (1999). How does financial pressure affect firms? *European Economic Review*, *43*(8), 1435–1456.

Odén, A., Wedel, H. Et al. (1975). Arguments for fisher's permutation test. *The Annals of Statistics*, *3*(2), 518–520.

OECD. (2017a). *Educational opportunity for all: Overcoming inequality throughout the life course*.

OECD. (2017b). *PISA 2015 Results (Volume IV): Students' Financial Literacy* (tech. rep.). OECD Publishing. https://doi.org/http://dx.doi.org/10.1787/9789264270282-en

OECD. (2017c). *PISA 2015 Technical Report* (tech. rep.).

Ololube, N. P. (2006). Teachers job satisfaction and motivation for school effectiveness: An assessment. *Essays in Education*, *18*, 1–10.

Ottaviano, G. I. (2011). *Firm heterogeneity, endogenous entry, and the business cycle* (tech. rep.). National Bureau of Economic Research.

Peek, J., & Rosengren, E. S. (2005). Unnatural selection: Perverse incentives and the misallocation of credit in japan. *American Economic Review*, *95*(4), 1144–1166.

Peng, R. D., & Matsui, E. (2015). The art of data science. *A Guide for Anyone Who Works with Data. Skybrude Consulting, LLC*.

Pesando, L. M. (2018). Education Economics Does financial literacy increase students' perceived value of schooling? Does financial literacy increase students' perceived value of schooling? *Education Economics*, *26*(5), 488–515. https://doi.org/10.1080/09645292.2018.1468872

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017). Psychological forest: Predicting human behavior, In *Thirty-first aaai conference on artificial intelligence*.

Psacharopoulos, G., & Patrinos, H. A. (2018). Returns to investment in education: A decennial review of the global literature. *Education Economics*, *26*(5), 445–458.

Rice, J. K. (2010). The impact of teacher experience: Examining the evidence and policy implications. brief no. 11. *National center for analysis of longitudinal data in education research*.

Riitsalu, L., & Põder, K. (2016). A glimpse of the complexity of factors that influence financial literacy. *International Journal of Consumer Studies*, *40*(6), 722–731. https://doi.org/10.1111/ijcs.12291

Robins, J. M., & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine*, *16*(3), 285–319.

Romer, P. M. (1989). *Human capital and growth: Theory and evidence* (tech. rep.). National Bureau of Economic Research.

Rosenbaum, P. R. (2002). *Observational studies*. New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, *45*(2), 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 34–58.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson Education Limited,

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210–229.

Schivardi, F., Sette, E., & Tabellini, G. (2017). Credit misallocation during the european financial crisis. *Bank of Italy Temi di Discussione Working Paper, No. 1139*.

Sejnowski, T. J. (2018). *The deep learning revolution*. MIT Press.

Shin, K.-S., Lee, T. S., & Kim, H.-j. (2005). An application of support vector machines in bankruptcy prediction model. *Expert systems with applications*, *28*(1), 127–135.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R., & Scott, J. G. (2018). Functional response regression with funbart: An analysis of patient-specific stillbirth risk. *arXiv preprint, arXiv:1805.07656*.

Starling, J. E., Murray, J. S., Lohr, P. A., Aiken, A. R., Carvalho, C. M., & Scott, J. G. (2019). Targeted smooth bayesian causal forests: An analysis of heterogeneous treatment effects for simultaneous versus interval medical abortion regimens over gestation. *arXiv preprint arXiv:1905.09405*.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*(4), 323.

Su, L., Shi, Z., & Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, *84*(6), 2215–2264.

Su, X., Kang, J., Fan, J., Levine, R. A., & Yan, X. (2012). Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research*, *13*(Oct), 2955–2994.

Sun, J., Fujita, H., Chen, P., & Li, H. (2017). Dynamic financial distress prediction with concept drift based on time weighting combined with adaboost support vector machine ensemble. *Knowledge-Based Systems*, *120*, 4–14.

Sun, J., & Li, H. (2011). Dynamic financial distress prediction using instance selection for the disposal of concept drift. *Expert Systems with Applications*, *38*(3), 2566–2576.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, *101*(476), 1619–1637.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tsai, C.-F., Hsu, Y.-F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, *24*, 977–984.

Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications*, *34*(4), 2639–2649.

Twala, B., Jones, M., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, *29*(7), 950–956.

Udo, G. (1993). Neural network performance on the bankruptcy classification problem. *Computers & industrial engineering*, *25*(1-4), 377–380.

Universiteit Gent. (2017). *Financiële geletterdheid bij 15-jarigen. Vlaams Rapport PISA 2015.* (tech. rep.).

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1).

Van Rijsbergen, C. J. (1979). Information retrieval.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

Vaughan, T. S., & Berry, K. E. (2005). Using monte carlo techniques to demonstrate the meaning and implications of multicollinearity. *Journal of Statistics Education*, *13*(1).

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.

Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, *41*(5), 2353–2361.

Wang, G., Li, J., & Hopp, W. J. (2018). An instrumental variable tree approach for detecting heterogeneous treatment effects in observational studies. *Available at SSRN 3045327*.

Wang, T., & Rudin, C. (2017). Causal rule sets for identifying subgroups with enhanced treatment effect. *arXiv preprint arXiv:1710.05426*.

Weinblat, J. (2018). Forecasting european high-growth firms-a random forest approach. *Journal of Industry, Competition and Trade*, *18*(3), 253–294.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, *37*(23), 3309–3324.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838.

White, T. K., Reiter, J. P., & Petrin, A. (2018). Imputation in us manufacturing data and its implications for productivity dispersion. *Review of Economics and Statistics*, *100*(3), 502–509.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, Massachusetts, MIT Press.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Nelson Education.

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, *7*(Nov), 2541–2563.

Zhao, Q., Small, D. S., & Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*(4), 735–761.

Zhou, L. (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. *Knowledge-Based Systems*, *41*, 16–25.

Zhou, Z., Athey, S., & Wager, S. (2018). Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*.

Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, *58*, 93–101.