Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics and Operational Research

Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Investment Readiness Programs for Start-Ups and SMEs in the Western Balkans

Luuk Vos (577988)

Supervisor:	Max Welz
Second assessor:	Prof. dr. Philip Hans Franses
Date final version:	1st July 2024

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics, or Erasmus University Rotterdam.

Abstract

This thesis explores the integration of Machine Learning (ML) techniques with Randomized Controlled Trials (RCTs) to analyze heterogeneous treatment effects, focusing on Investment Readiness Programs (IRPs) for start-ups and SMEs in the Western Balkans. Traditional RCT methods primarily evaluate the average treatment effect (ATE), which often overlook possible heterogeneity. This study aims to address these limitations by implementing ML, providing a better understanding of how various groups respond to interventions. RCTs are frequently used for minimizing biases and enabling causal inference but are often costly, time-consuming, and sometimes even ethically constrained. Contrariwise, ML methods excel in handling large datasets and uncovering complex patterns, which make them suitable for detecting treatment effect heterogeneity. However, obtaining consistent estimators for conditional average treatment effects (CATE) using ML techniques can be challenging due to high-dimensional settings and the need for additional assumptions. The paper employs a generic ML approach to estimate key features of the CATE, enhancing the robustness and flexibility of non-parametric inference and avoiding consistency issues. These are the Best Linear Predictor (BLP), Sorted Group Average Treatment Effects (GATES), and Classification Analysis (CLAN). Furthermore, this paper's results indicate a significant finding: firms with 5-6 employees at the start of the process experience significantly less treatment effect, primarily driven by those with exactly 6 employees. This deviates from the original conclusions, suggesting that traditional methods might have overlooked this detail. Additionally, the generic ML method does not perform well with lower-dimensional settings, which indicates that simpler linear regression techniques might have sufficed.

1 Introduction

Randomized Controlled Trials (RCTs) and Machine Learning (ML) methods are powerful approaches in modern research and data analysis. RCTs are considered the best standard for evaluating the efficiency of interventions. Namely, the method randomly assigns the participants to either the *treatment* or the *control* group. Thereby, it minimizes biases, such as selection and performance bias, and allows for causal inference. However, RCTs can be costly and time-consuming. In some cases, it can even face ethical constraints, such as not being able to withhold potentially beneficial treatments from the control group (Levine, 1981). The main focus of traditional methods usually is on the average treatment effect (ATE), without considering treatment effect heterogeneity among various subgroups within the population. Some of these analysis methods include t-tests, ANOVA, and regression analysis to help identify the ATE.

ML methods can enhance the analysis of RCTs by addressing these shortcomings. These methods offer complex problems to be solved using, often advanced, computational techniques. Specifically, the methods are superior in working with large and intricate datasets, uncovering patterns, and forecasting. Hence, in the context of causal learning, they can also help detect treatment effect heterogeneity and provide insights into how different subgroups respond to the intervention. Consequently, more precise and personalized treatment recommendations can be given. This paper focuses on integrating these two methodologies to further enhance Investment Readiness Programs (IRPs).

To better understand the methods used, the following concepts are clarified in the context of RCTs and ML applications: treatment effects, causal inference, and non-parametric inference. Starting with treatment effects in RCTs, the effects are defined as a measure implicating the differences in outcomes between the *treatment* and the *control* group. These effects are typically presented using the mean difference, known as the Average Treatment Effect (ATE). It is self-evident that estimating this effect accurately is crucial for evaluating the outcomes of an intervention. In this paper, heterogeneity in treatment effects is specifically discussed. Hence, the examination is focused on how these effects vary across various subgroups.

Next, causal inference refers to determining whether there is a cause-and-effect relationship between the subgroups and variables. Thereby, it ensures that the observed changes can be accredited to the *treatment* instead of other external factors. However, finding these relationships can be complex, especially in the case of high-dimensional variables. In the context of this paper, ML techniques can be useful specifically in treatment effect heterogeneity due to their ability to capture these intricate interactions.

Finally, non-parametric inference offers RCTs greater flexibility and robustness. The non-parametric methods can estimate treatment effects without relying on traditional parametric assumptions, such as linearity, normality, or other distributional forms of data. As a result, the flexibility increases the accuracy of modeling complex relationships. However, in turn, it introduces difficulties in estimation and inference. Once again, the main issue lies in high-dimensional settings. In such cases, traditional inference techniques regularly do not produce consistent estimators for conditional average treatment effects (CATE), which represents the difference in expected potential outcomes between treated and control groups, conditional on covariates. In Chernozhukov et al. (2023), the authors propose utilizing ML techniques to create proxy predictors for CATE, knowing that ML methods manage to handle high-dimensional data and find complex patterns adequately. As a result, the estimation and inference are valid.

This paper makes use of a generic ML approach, meaning it can be applied in conjunction with any ML method, introduced by the authors in Chernozhukov et al. (2023). Their method addresses the fundamental challenges in non-parametric inference, particularly the difficulties of applying ML tools to estimate heterogeneous causal effects. Motivated by the work of Genovese & Wasserman (2007), this approach focuses on achieving valid estimation and inference on the features of the CATE. In Chernozhukov et al. (2023), they start with an ML proxy predictor of the CATE, which is consequently used as inspiration to target its features. The first feature is the *Best Linear Predictor* (BLP) of the CATE on the proxy predictor, allowing the determination of which method fits the data best. Secondly, the *Sorted Group Average Treatment Effects* (GATES), which represent the average treatment effect for groups categorized by the ML proxy predictor based on heterogeneity. Lastly, *Classification Analysis* (CLAN) examines the average characteristics of the most and least affected units as determined by the ML proxy predictor. Hence, the features allow detection of observable heterogeneity, determine the treatment effect for different segments, and identify which covariates correspond to the heterogeneity.

The authors further propose quantile-aggregated inference, which addresses the variability introduced by different sample splits by aggregating results. This approach involves taking medians of estimates and other quantiles from different splits. By doing so, quantile aggregation reduces the likelihood that two researchers working with the same data will draw different conclusions. It also lowers estimation risks compared to a single-split procedure and its inferential properties are well-established.

Finally, to compare and select among ML methods, they develop goodness-of-fit measures for the BLP and GATES. These measures are used to construct ML proxies that more accurately target the CATE through causal learning. Consequently, these causal models produce better CATE proxies than generic predictive ML methods. This approach allows the post-processing methods to focus on providing valid inferences without needing to correct for biases.

This paper steps away from the microfinancing application in Chernozhukov et al. (2017) and applies RCTs and the corresponding ML methods to venture capital and entrepreneurship. More specifically, the focus lies on start-ups and small and medium-sized enterprises (SMEs), and the impact of IRPs in the Western Balkans. In Cusolito et al. (2021), the authors evaluate the investment readiness component of the Western Balkans Enterprise Development and Innovation Facility (WB EDIF). According to WBIF (2024), the initiative is financed by the European Commission to mainly improve the access to finance for SMEs in the region. This is the result of the embarrassing statistic from Vizjak & Vizjak (2016) that in 2014 only 21.8 million euros were received by 15 start-ups from venture capital funds and business angels in Croatia. Additionally, Gattini et al. (2016) reported at most two cases of investments in both Kosovo and Montenegro.

Furthermore, it is notable that Croatia has decreased its R&D expenditure between 2000 to 2018, from 1.04% to 0.95% (World Bank, 2024). This remains low, especially when compared to the two of the most thriving economies: China and the US. According to Mallaby (2023), China increased its R&D share of its rapidly growing GDP from 0.9% to 2.1%, during the same period, whereas the US remained at its high percentile going from 2.5% to 2.8%. This inevitably proves the importance of R&D expenditure to the economy.

Start-ups and SMEs play a crucial role in economic growth and innovation, especially being the driving forces behind new technologies and job creation. However, securing funding and resources for these firms to succeed is generally challenging. The study mentioned above of Cusolito et al. (2021) recognizes that this is especially true for smaller firms stationed in the Western Balkans. They find that these developing companies often have not fine-tuned their ideas to the stage where they can attract outside funding or are unwilling to surrender any ownership stake, referring to Mason & Kwok (2010). Start-up accelerators have proven to significantly help in their journeys, mentoring the firms to make the right decisions. In Del Sarto et al. (2022), the author states that an accelerator's highly specialized network alone already manifests to be highly effective at transferring knowledge to improve the innovation performance of start-up teams. However, these accelerator companies generally work with small cohorts of 10 to 20 firms at a time and are expensive. Hence, IRPs have been a solution to this issue, providing a less expensive, yet similar mentoring approach that can be scaled to support a larger number of start-ups and SMEs simultaneously.

IRPs are intended to provide a comprehensive approach to overcoming the constraints that firms face in receiving outside investment. They combine individualized training, mentoring, and coaching to make firms more investment-ready. Thereby, assisting the firms in attracting funding by fine-tuning their business models, improving their pitches, and strengthening their financial planning. In an attempt to provide causal evidence, Cusolito et al. (2021) conducted a five-country RCT in Croatia, Kosovo, Macedonia, Montenegro, and Serbia. The treatment group participates in an IRP, whereas the control group receives a basic online investment readiness course. As Cusolito et al. (2021) explain, providing the control group with a form of treatment instead of no treatment minimizes the risk of Hawthorne and John Henry effects. Subsequently, the firms participate in a pitch event, where judges determine their investment readiness scores. After 6 months, the firms are asked to fill in a follow-up survey, determining whether serious steps and/or improvements have been made. However, the main focus is on the results of a second follow-up survey, conducted approximately two years after completing the program for the same reasons. In this paper, the generic ML method is applied to this dataset using Welz et al. (2022), led by the following research question: *How does the Investment Readiness Program (IRP) impact access to equity finance for innovative start-ups and small and medium-sized enterprises (SMEs) in the Western Balkans, and how do the heterogeneous treatment effects vary across different firm characteristics, using machine learning and causal inference methods?*

Applying machine learning methods to this dataset can uncover nuanced patterns and heterogeneity in treatment effects that the linear regression in Cusolito et al. (2021) might have missed. Identifying subgroups and higher precision of treatment effect estimation is particularly useful for equity finance. These improvements can provide new insights into which elements of the IRP are most effective, allowing more targeted and effective intervention strategies to be developed. Therefore, this study aims to further investigate the relationships Cusolito et al. (2021) claim to have found.

Whilst replicating the results in Cusolito et al. (2021), the conclusions do not replicate under a different grouping, suggesting they may not be robust to sensitivity analysis. Consequently, this provides motivation to conduct a more comprehensive investigation, by employing another grouping of the company sizes that were noted when registering. This ensures the robustness of the observed relationship. Due to this different grouping, the theory is that the original conclusion is based on a dip caused by the 6-employee sized firms, instead of the IRP having a lower effect on larger firms in general. This study supports this theory with some certainty. From the results, it is determined that only the firms with 5 to 6 employees have a significantly lower treatment effect than other groups, suggesting heterogeneity in the company sizes. Furthermore, the ML models provided wide 90% confidence intervals on the insignificant estimates, suggesting that the dataset is too small for robust and reliable results.

The remainder of the paper is structured as follows. Section 2 formalizes the key features of the CATE, discusses the goodness-of-fit measures, and the use of causal machines. Section 3 presents the key characteristics of the data and describes how it is enhanced through imputation. Section 4 compares the grouping in the replication with that of the current study and further analysis. Section 5 concludes with some remarks. Furthermore, the appendices contain the replication results, detailed descriptions of parts of the methodology, data, and results, and how to use the replication code.

2 Methodology

In this section, the methodology of the generic machine learning (ML) approach is outlined by considering the data as $(Y_i, Z_i, D_i)_{i=1}^N$, which comprises independent and identically distributed (i.i.d.) samples of the random vector (Y, Z, D). Here, Y represents the outcome variable, D is a binary treatment indicator, and Z is a potentially high-dimensional vector of covariates describing the observational units.

In this paper, the approach is the same as in the study conducted by Chernozhukov et al. (2023), where the authors propose unique strategies for tackling estimation and inference challenges. Namely, the paper is constructed on strategies on *key features* rather than on the target function itself. This function is the conditional average treatment effect (CATE):

$$s_0(Z) := E[Y(1) - Y(0)|Z] = E[Y(1)|Z] - E[Y(0)|Z], \tag{1}$$

where $Y(1)$ and $Y(0)$ are the potential outcomes in the treatment state 1 and the non-treatment state 0. The CATE is one of two main causal functions, the other is the baseline conditional average (BCA):

$$b_0(Z) := E[Y(0)|Z]. \tag{2}$$

As can be observed from the expected values being conditional on the covariates in Equation 1, $s_0(Z)$ is complex, logically making its estimation and inference challenging. ML methods are not an exception as they often rely on sparsity assumptions. These assumptions make the function dependent only on a small subset of covariates. However, taking a subset clashes with the need for assumptions as selecting a small subset implies that these must be the only relevant covariates for estimation. Yet, without such assumptions, it is difficult, if not impossible, to obtain consistent estimators for $s_0(Z)$. Moreover, the minimax lower bound implies that learning $s_0(Z)$ from a finite sample size N cannot be achieved at a reasonable rate as the dimensionality increases. As a result, ML estimators can fail to produce consistent estimates in high-dimensional settings unless additional, often untestable, assumptions are made. Hence, by focusing on the features instead, the authors neglect these challenges of obtaining consistent estimators of the target function. They do this by treating ML as providing proxy predictors for the key features of $s_0(Z)$ and by relying on sample splitting in the process.

A random partition (M, A) of the set of indices $\{1, \ldots, N\}$ results in a main sample M, $Data_M = (Y_i, D_i, Z_i)_{i \in M}$, and an auxiliary sample A, $Data_A = (Y_i, D_i, Z_i)_{i \in A}$. After splitting the sample, two stages are followed. *Stage 1* uses A to obtain (not necessarily consistent) ML estimators of the baseline functions and treatment effects. These are named the ML proxy predictors,

$$z \mapsto B(z) \text{ and } z \mapsto S(z),$$

where both $B(Z)$ and $S(Z)$ are possibly biased and noisy predictors of $b_0(Z)$ and $s_0(Z)$, respectively. Please refer to Section 2.3 for the implementation of causal machines in this stage. In *stage 2*, the proxies from stage 1 are post-processed to estimate and make inferences on, in this paper, three features of the CATE function $z \mapsto S(z)$. These key features are the *Best Linear Predictor* (BLP), *Sorted Group Average Treatment Effects* (GATES), and *Classification Analysis* (CLAN), which are thoroughly discussed in Section 2.1.

To tackle the uncertainty that comes with sample splitting, a high number q of sample splits is completed that allows for quantile aggregation of inference to combine results across the splits. This procedure uses the median of the estimations over the q splits to give point estimation. Similarly, for interval estimation and p-values, the medians of conditional confidence sets and p-values are used, respectively.

Please note that the definitions used in this section are similar or the same as those used in Chernozhukov et al. (2023). Furthermore, the theorems and their corresponding referential numbers are those from the original paper.

2.1 Key Features of the CATE

In defining the key features, the functions are conditioned on $Data_A$, resulting in the following fixed functions

$$z \mapsto B_A(z) := B(z; Data_A) \text{ and } z \mapsto S_A(z) := S(z; Data_A).$$

Similarly, $E[\cdot|Data_A]$ is referred to by using $E_A[\cdot]$.

Moreover, the first two subsections cover BLP and GATES, where Chernozhukov et al. (2023) thoroughly discuss two strategies for identifying and estimating the BLP and GATES. The first, *Strategy A*, uses a weighted linear projection approach, whereas *Strategy B* makes use of the Horvitz-Thompson transform H variable. The variable is the residualized treatment scaled by its variance, and is given by

$$H = H(D, Z) := \frac{D - p(Z)}{p(Z)(1 - p(Z))}.$$

Theoretically, *Strategy A* should outperform its counterpart due to its stability, simplicity, and noise management. Namely, the HT transformation can give a noisy signal, which in turn can be mitigated by additional methods like including baseline covariates or using residualized outcomes, but requires larger datasets. Hence, *Strategy A* should perform slightly better in this case, due to the lower dimensional dataset. Nonetheless, in comparison, both strategies lead to the same conclusions, with the HT transformation giving slightly better results. Hence, this paper solely focuses on *Strategy B*.

2.1.1 Best Linear Predictor of CATE (BLP)

An important inferential target is identifying the BLP using the proxy $S(Z)$. The following definition provides a formal explanation, explicitly showing the dependence on Data_A.

Definition 1 (BLP). The best linear predictor of $s_0(Z)$ by $S_A(Z)$ is the solution to:

$$\min_{b_1, b_2} E_A[s_0(Z) - b_1 - b_2 S_A(Z)]^2,$$

which, if it exists, is defined as

$$\text{BLP}[s_0(Z) \mid S_A(Z)] := \beta_1 + \beta_2(S_A(Z) - E_A[S_A(Z)]),$$

where $\beta_1 = E_A[s_0(Z)]$ and $\beta_2 = \frac{\text{Cov}_A[s_0(Z), S_A(Z)]}{\text{Var}_A[S_A(Z)]}$.

As can be observed, the BLP uses $S_A(Z)$ to become an unbiased predictor of $s_0(Z)$ by improving upon $S_A(Z)$ in terms of the mean-squared error. This improvement is quantified by showing that the mean-squared error of $S_A(Z)$ minus the BLP is always positive unless $S_A(Z)$ is already an unbiased predictor and $\beta_2 = 1$ or $\text{Var}[S_A(Z)] = 0$. Therefore, the BLP can be seen as a refined predictor of $s_0(Z)$. Generally, $\beta_2 \neq 1$ to correct for noise in $S_A(Z)$. However, if $\beta_2 = 1$, then there is no noise, and $S_A(Z)$ is considered a perfect proxy for $s_0(Z)$. Applying the same logic, if $\beta_2 = 0$, then $S_A(Z)$ is complete noise and it can be concluded there is no heterogeneity, as $s_0(Z)$ is linked to $B_2 = 0$. Hence, rejecting the hypothesis $B_2 = 0$ allows the conclusion that there is heterogeneity in $s_0(Z)$ and $S_A(Z)$ is a relevant predictor.

As mentioned, *Strategy B* makes use of the H to obtain the transformed response YH that provides an unbiased signal, ensuring that $E_A[YH|Z] = s_0(Z)$. Hence, using Definition 1, it follows that the BLP can be written as

$$\text{BLP}[s_0(Z)|S_A(Z)] = \text{BLP}[YH|S_A(Z)].$$

From here, two key points should be noted. First, the BLP is more precise than YH, as the variance of the BLP is less than or equal to the variance of YH, by the law of total variance. Second, its simple linear projection $\text{BLP}[YH|S_A(Z)]$ can underperform in estimation and inference due to lack of precision. Consequently, the following linear projection of YH is considered

$$YH = \mu_0' X_1 H + \mu_1 + \mu_2(S_A(Z) - E_A[S_A(Z)]) + \epsilon, \quad E_A[\tilde{X}] = 0, \tag{3}$$

where $\tilde{X} := (X_1'H, \tilde{X}_2)'$, $\tilde{X}_2 := (1, S_A(Z) - E_A[S_A(Z)])'$, and $X_1 := [1, B(Z), p(Z), p(Z)S_A(Z)]'$. Please note that X_1 could contain other functions of Z, and X_1H included to reduce noise.

Using Theorem 3.2 (Theorem 1 in Appendix B.1), Equation 3 can be estimated using specific linear regression, which refers to applying linear regression techniques tailored to the structure and requirements of this model to ensure accurate estimation. This estimation is given in Equation 10 in Appendix A.1.

2.1.2 Sorted Group Average Treatment Effects (GATES)

The second goal of this inferential analysis is to determine the set of group average treatment effects, which are defined by the groups induced by $S_A(Z)$. Logically, the groups are designed to explain as much variation in $s_0(Z)$ as possible. Similarly to BLP, the following definition is provided in Definition 2.

Definition 2 (GATES). The Sorted Group Average Treatment Effects (GATES) are

$$\gamma_k := E_A[s_0(Z)|G_k], \quad k = 1, \ldots, K.$$

where $G_k := \{S_A(Z) \in I_k\}$, with $I_k := [\ell_{k-1}, \ell_k)$ and $-\infty = \ell_0 < \ell_1 < \ldots < \ell_K = +\infty$.

Furthermore, *Strategy B* employs a linear projection on Horvitz-Thompson transformed variables to identify the GATES. The linear projection equation that is used is

$$YH = \mu_0 X_1 H + \sum_{k=1}^{K} \mu_k \cdot 1(G_k) + \nu, \quad E_A[\nu \tilde{W}] = 0, \tag{4}$$

where $\tilde{W} := (X_1'H, \tilde{W}_2')'$, X_1 includes functions of Z, e.g., X_1 the same as above, and $\tilde{W}_2 := \{1(G_k)\}_{k=1}^{K}$.

Once again, Equation 4 can be estimated using specific linear regression, by Theorem 3.3 (Theorem 2 in Appendix B.2), resulting in Equation 11 in Appendix A.2.

2.1.3 Classification Analysis (CLAN)

The last key feature of the CATE is the CLAN, which allows finding properties of various groups that are most and least affected by the treatment. Whilst the results of the analysis are most interesting when there is substantial heterogeneity in the samples, the analysis supports finding heterogeneity in the K subgroups regardless of whether BLP and GATES analyses have revealed heterogeneity in the first place. The CLAN specifically focuses on quantifying the differences between the groups, while providing the responsible covariates in the process. Furthermore, let G_1 and G_K denote the *least affected group* and *most affected group*, respectively, where the labels of *most* and *least* might differ depending on the context. In Chernozhukov et al. (2023), the authors give their definition in Definition 3.

Definition 3 (CLAN). Let $g(Y, D, Z)$ be a vector of characteristics of an observational unit. The classification analysis (CLAN) is the comparison of the average characteristics of the most and least affected groups:

$$\delta_1 := E_A[g(Y,D,Z) \mid G_1] \quad \text{and} \quad \delta_K := E_A[g(Y,D,Z) \mid G_K].$$

In Definition 3, the parameters δ_1 and δ_K are averages of variables that are directly observed. Hence, they are identified. Moreover, it is important to note that the CLAN is not merely fixed on comparisons on averages, but can also be shifted to focus on variances, covariances, or distributions. However, in this paper, this extension is not utilized.

δ_1 and δ_K are estimated by taking averages in M:

$$\hat{\delta}_1 = \frac{\mathbb{E}_{N,M}[g(Y_i,D_i,Z_i)G_{1,i}]}{\mathbb{E}_{N,M}[G_{1,i}]} \quad \text{and} \quad \hat{\delta}_K = \frac{\mathbb{E}_{N,M}[g(Y_i,D_i,Z_i)G_{K,i}]}{\mathbb{E}_{N,M}[G_{K,i}]}, \tag{5}$$

using $G_{k,i} = 1\{S(Z_i) \in I_k\}$, where $I_k = [\ell_{k-1}, \ell_k)$ and ℓ_k is the (k/K)-quantile of $\{S_A(Z)_i\}_{i \in M}$. Furthermore, the operation $\mathbb{E}_{N,M}$ refers to the empirical expectation concerning the main sample M.

2.2 Goodness of Fit Measures for Fitting CATE

In this subsection, the goodness-of-fit measures that guide the selection of the ML proxies are introduced. Starting with the BLP of CATE, the following definition is given:

$$\Lambda := |\beta_2|^2 \text{Var}(S_A(Z)) = \text{Corr}(s_0(Z), S_A(Z))^2 \text{Var}(s_0(Z)). \tag{6}$$

The goal is to maximize Λ in obtaining the preferred method. This can be observed by noting that maximizing Λ corresponds to maximizing the correlation between the ML proxy predictor $S_A(Z)$ and CATE $s_0(Z)$. Equivalently, it is maximizing the R^2 value in the regression of $s_0(Z)$ on $S_A(K)$. Hence, maximization is clearly more desirable.

Similarly, the following is proposed for the GATES analysis:

$$\bar{\Lambda} = E_A \left(\sum_{k=1}^{K} \gamma_k 1\{S_A(Z) \in I_k\} \right)^2 = \sum_{k=1}^{K} \gamma_k^2 P(S_A(Z) \in I_k). \tag{7}$$

Here, the goal is to maximize $\bar{\Lambda}$, as it is equivalent to maximizing the R^2 in the regression of $s_0(Z)$ on $\bar{S}(Z)$ (without a constant). This can be observed by noting that in Equation 7, $\bar{\Lambda}$ represents the part of variation of $s_0(Z)$, $E_A[s_o(Z)^2]$, explained by the mean of the ML proxy, $\bar{S}_A(Z) = \Sigma_{k=1}^{K} \gamma_k 1\{S_A(Z) \in I_k\}$. If the groups $G_k = \{S \in I_k\}$ have equal sizes, implying $P(S(Z) \in I_k) = 1/K$ for each $k = 1, \ldots, K$, then $\bar{\Lambda} = \frac{1}{K}\sum_{k=1}^{K} \gamma_k^2$. Hence, maximizing $\bar{\Lambda}$ is preferred. Empirically, the parameters are defined as

$$\hat{\Lambda} = |\hat{\beta}_2|^2\, \mathbb{E}_{N,M}(S_A(Z)_i - \mathbb{E}_{N,M}\, S_A(Z)_i)^2, \quad \hat{\bar{\Lambda}} = \sum_{k=1}^{K} \hat{\gamma}_k^2\, \mathbb{E}_{N,M}\, 1\{S_A(Z)_i \in I_k\}. \tag{8}$$

In the method selection, two best-performing ML methods are chosen based on these Λ and $\bar{\Lambda}$ values. If there is one model that dominates both metrics, then the best method is given. Yet, if multiple methods perform similarly, different approaches can be used, as implied by Chernozhukov et al. (2023) in Appendix A.3. Nonetheless, the focus of this paper lies in identifying heterogeneity in the treatment effect rather than determining the optimal ML method. It can be argued that these go hand in hand. However, as long as the differences in the metrics between the methods are minor, the priority is on the identification of treatment effect heterogeneity. This approach provides more immediate insights into the causal relationships within the data, whilst conserving resources.

2.3 Estimating CATE through Causal Machines in First Stage

This section uses causal machines to target CATE directly in the first stage. In Chernozhukov et al. (2023), two options are proposed similarly to the strategies A and B in Section 2.1. *Option A* refers to minimizing $w(Z)$-weighted square prediction errors of Y on B and $(D - p(Z))$S, whereas *Option B* considers minimizing square prediction errors of YH on BH and S. Here, B refers to $B(Z)$ that is considered a "baseline" function of covariates Z. The covariates have a role in reducing noise in the learning problem. Similarly to Section 2.1, the preference goes towards *Option B* due to the Horvitz-Thompson transform H providing slightly better results. Consequently, the causal learners for stage 1 are defined as follows:

Definition 4 (Causal Learners for Stage 1). The HT learner is solved:

$$(B, S) \in \arg \min_{b \in \mathcal{B}, s \in \mathcal{S}} \sum_{i \in A} [Y_i H_i - b(Z_i) H_i - s(Z_i)]^2 \tag{9}$$

where $w(Z) = [p(Z)(1 - p(Z))]^{-1}$, and $\mathcal{B}$ and $\mathcal{S}$ are functional parameter spaces.

Some examples of parameter spaces that Equation 9 refers to can include spaces of linear functions generated by a set of dictionary transformations of Z, reproducing kernels, linear combinations of decision trees, neural networks, and other forms. It is important to note that these parameters spaces are intended, though not required, to include $z \mapsto \bar{b}_0(Z) := b_0(Z) + (1 - p(z))s_0(z)$ and $z \mapsto s_0(Z)$.

Furthermore, some attractive theoretical properties arise from this approach. Specifically, these properties demonstrate that causal learners can closely approximate the true CATE function even without direct observation. These are named the oracle properties, which go beyond the scope of this paper. For the interested reader, these are thoroughly discussed in the context of this paper in Appendix A.5.

2.4 Implementation

In Algorithm 1, the details of the methods and steps taken in the heterogeneity analysis are presented. As can be observed, the algorithm median-aggregates the p-values and the point and interval estimators. Thereby, it accounts for the uncertainty arising from parameter estimation and sample splitting, ensuring the validity and reliability of the results. These calculations are further explained with precise detail in Appendix A.4.1. Furthermore, three technical conditions are implemented that improve the consistency of the median-aggregated p-values and the properties of the confidence intervals for the target parameter, as detailed in Appendix A.4.2.

The algorithm and conditions are implemented using Welz et al. (2022), which is Max Welz's *GenericML* package in R.

```
Algorithm 1: Inference Algorithm
1 Inputs: The inputs are given by the data $\{(Y_i, D_i, Z_i, p(Z_i))\}$ on units
    $i \in [N] = \{1, \ldots, N\}$. Fix the number of splits $N_S$ and the significance level $\alpha$, e.g.,
    $N_S = 100$ and $\alpha = 0.05$. Fix a set of ML or Causal ML methods.
    1. Generate $N_S$ random splits of $[N]$ into the main sample, $M$, and the auxiliary
       sample, $A$. Over each split apply the following steps:
        (a) Using $A$, train each ML method and output predictions $B_A(Z)$ and $S_A(Z)$ for $M$.
        (b) Optionally, choose the best or aggregate ML methods using the results of
            Section 2.3.
        (c) Estimate the BLP parameters via HT BLP (Equation 10) in $M$.
        (d) Estimate the GATES parameters by HT GATES (Equation 11) in $M$.
        (e) Estimate the CLAN parameters by taking averages (Equation 5) in $M$.
        (f) Compute the goodness of fit measures via (Equation 8) in $M$.
    2. If the winning ML methods were not chosen in Step 1b, median-aggregate the
       goodness-of-fit measures and choose the best ML methods.
    3. Compute and report the quantile-aggregated point estimates, p-values, and
       confidence intervals of Appendix A.4. If Step 2 is used, compute and report the union
       of these statistics for all winners.
```

3 Data

The data used in this paper is obtained from Cusolito et al. (2021). In their paper, a five-country RCT was set up to evaluate the gain in participation in Investment Readiness Programs (IRPs) for start-ups and SMEs. From the application onwards, the participants had to answer survey questions about their company's past, present, and future. Consequently, the applications eligible for the program were then assessed for their initial investment readiness level. This assessment was based on four criteria: market attractiveness, product technology, traction, and team.

The top 10 firms, based on their initial scores, were evenly split into the treatment and control groups. Furthermore, the remaining firms were divided into strata based on the country registered and whether they already had a private investor. Moreover, the innovative firms were ranked into quartets within these strata based on their initial scores. These quartets would then be split into the treatment and control groups.

The treatment group participates in an IRP, an extensive program containing individualized training, mentoring, and coaching to make firms more investment-ready. Subsequently, the control group takes part in a simple online investment readiness course, in contrast to regular control groups not receiving any treatment at all. This unorthodox allocation rather shifts the focus to whether expensive IRPs are worth it. Furthermore, offering both groups a program is beneficial in lowering the risk of Hawthorne and John Henry effects, as explained in Cusolito et al. (2021). To define, the Hawthorne effect is when the treatment group changes behavior due to the existence of observation, whereas John Henry is when that is the case for the control group. Delving deeper into the programs, the difference between one-on-one and online treatment lies mainly in the focus, attention, and intensity. For instance, a difference lies in that the high-cost and intensive program involved receiving a mentor to help develop financial plans, product pitches, market strategy, and willingness to take equity financing.

The initial randomization process involved 333 firms, with 167 assigned to the treatment group and 166 to the control group. Later, the firms that took longer to verify were added by creating separate strata based on investment scores. This resulted in 346 firms, with 174 in the treatment group and 172 in the control group. Unfortunately, not all firms attended the semi-finals, reducing the firms eligible for this experiment to 211. Nonetheless, the treatment and control groups were still in balance with 110 and 101 firms attending, respectively. In Table 3.1, the descriptive statistics along with their balance test p-values are presented for both the initial batch and the semi-final attendees. Notably, the statistics between the initial group and semi-finalists are extremely similar, suggesting the strata are still in place for the semi-finalists after the major cut in participants. Furthermore, all treatment and control group pairs are well balanced, as observed from the p-values. However, this is notable considering the insignificant large difference of 103,360 for *revenue in 2014* for the semi-finalist group. The groups still being in balance suggest wide confidence intervals, which could be an indicator that the dataset is too small.

Furthermore, Cusolito et al. (2021) measured various outcomes across the different firms, such as media buzz, having acquired an investor, or even the general need to find external financing. Outcomes like the last two examples are based on the follow-up surveys by the semifinalists, which had high response rates. The first follow-up survey was conducted approximately six months after the end of the program and competition, which had responses from 79% of firms. Approximately 2 years after the end of the program, the second follow-up survey had an 85% response rate. Specifically, these surveys focused on measuring changes in three domains. The first is whether the firm is still operating, the second is investment readiness, and the third looks at the steps taken toward receiving external funding or external financing already received.

Investment readiness focuses on three aspects, according to Mason & Kwok (2010). The first is that the owners need to have significant willingness and interest in taking on equity. Second is whether the firm is a viable business of interest to investors, looking at employment, sales, and profits. This is measured as general investability. The final aspect is the collection of specific measures that investors require for an investment, such as "separation of outcomes, revenue projections, knowledge of customer acquisition costs, tracking key metrics of traction, and covering intellectual property", as summarized in Cusolito et al. (2021).

Even though the initial dataset contains 346 observations with 755 covariates, the dataset worked with diminishes to 211 observations and 69 covariates. The main reason for this being is that the surveys include follow-up questions that have been compromised to index values. Furthermore, the working dataset still contains 356 (2.9% of the data points) NA values that result from (part of) the surveys not being answered. If these rows are removed, the 211

Table 3.1: Selected Descriptive Statistics of the Firms

	Full Sample			Semi-Final Participants		
	Treatment	Control	p-value	Treatment	Control	p-value
Variables stratified on						
Incorporated/Registered in Croatia (Binary)	0.25	0.24	0.612	0.26	0.30	0.920
Incorporated/Registered in Serbia (Binary)	0.46	0.46	0.626	0.48	0.48	0.513
Baseline Readiness Score	2.95	2.92	0.150	2.99	2.97	0.476
Has an Outside Private Investor (Binary)	0.10	0.09	0.178	0.14	0.06	0.170
Other variables						
Market Attractiveness Score	3.08	3.05	0.851	3.13	3.18	0.579
Product Technology Score	2.47	2.43	0.835	2.56	2.71	0.085
Traction Score	3.34	3.27	0.507	3.28	3.06	0.382
Team Score	3.04	3.05	0.878	3.08	3.02	0.207
Sector is Business and Productivity (Binary)	0.48	0.39	0.107	0.45	0.36	0.436
Sector is Lifestyle and Entertainment (Binary)	0.18	0.23	0.295	0.20	0.27	0.215
Sector is Energy and Utilities (Binary)	0.07	0.06	0.816	0.07	0.07	0.506
Sector is Financial Services (Binary)	0.01	0.03	0.257	0.01	0.02	0.180
Sector is Life Science and Agriculture (Binary)	0.05	0.10	0.102	0.05	0.12	0.063
Sector is Materials and Manufacturing (Binary)	0.11	0.13	0.575	0.10	0.09	0.885
Sector is Mobility and Transporation (Binary)	0.10	0.06	0.251	0.12	0.08	0.055
Industry is Communication & Collaboration (Binary)	0.16	0.11	0.224	0.16	0.10	0.122
Industry is Marketing & Advertising (Binary)	0.18	0.09	0.012	0.18	0.06	0.021
Industry is IT (Binary)	0.11	0.10	0.818	0.11	0.09	0.679
Place in Value Chain is Developer (Binary)	0.61	0.55	0.171	0.60	0.57	0.677
Place in Value Chain is Researcher (Binary)	0.32	0.29	0.510	0.33	0.33	0.782
Place in Value Chain is Service Provider (Binary)	0.59	0.54	0.372	0.60	0.54	0.108
Place in Value Chain is Producer (Binary)	0.40	0.39	0.867	0.34	0.39	0.482
Place in Value Chain is Merchant (Binary)	0.12	0.14	0.589	0.13	0.15	0.838
Age of Firm (years)	2.61	2.66	0.887	2.24	2.29	0.346
Early Stage Firm (Binary)	0.30	0.33	0.475	0.35	0.37	0.554
Revenue in 2014	178,100	184,800	0.959	37,640	144,000	0.303
Number of Employees	6.47	5.88	0.539	4.65	5.32	0.800
Company has More than 4 Employees (Binary)	0.55	0.49	0.259	0.56	0.44	0.082
Age of Main Founder	38.22	36.81	0.204	38.02	36.67	0.362
Main Founder has Post-Graduate Education (Binary)	0.49	0.48	0.816	0.54	0.55	0.740
At Least One Founder is Female (Binary)	0.16	0.22	0.128	0.16	0.30	0.071
Company has a Global Focus (Binary)	0.60	0.58	0.576	0.59	0.63	0.569
Have Accepted Outside Funding (Binary)	0.34	0.37	0.656	0.42	0.40	0.836
Previously in Mentoring/Accelerator Programme (Binary)	0.15	0.16	0.704	0.18	0.22	0.202
Sample Size	174	172		110	101	

Notes: The treatment is participation in an IRP, where the control group participates in an online investment readiness course. The p-value refers to the balance test on the values for each variable. The *score* variables represent the baseline scores given by the judges before the competition and treatment. Moreover, *incorporated/registered* variables give the location that the firm is located. *Has an outside private investor* gives whether the firm already has an outside investor before participating. *Sector, industry*, and *place in value* variables mention in which category the firm falls into. *Early stage firm* is a binary variable being 1 if the firm has just started. *Company has more than 4 employees* indicates whether the firm has above the median number of employees.

observations (firms) are further decreased to 131. Consequently, the values are filled using kNN imputation, as is clarified in Appendix C.1.

4 Results

Before moving to the results of the heterogeneity analysis, the replication of the results presented in both Chernozhukov et al. (2017) and Cusolito et al. (2021) was first completed, given in Appendix E and Appendix F, respectively. Notably, in Cusolito et al. (2021), the authors conclude that in "examining the heterogeneity of impacts, the program appears to have only succeeded in increasing investment readiness and the chance of subsequent external financing for smaller firms (those with 1 to 3 workers), and those which otherwise were less likely to receive external financing." Hence, smaller firms supposedly received more gain from treatment than larger firms (4+). It is important to note that when referred to company size in this section, it represents the initial number of employees at the time of registration. Moreover, the authors visualize and quantify this finding by referring to Figure F and the significant values in Table F.4 before the conclusion. Whilst replicating Figure F, it was odd for the boundaries to be irregularly divided and double, as sample sizes of the various numbers of employees are similar and each size of the firm was incorporated twice. Changing these bounds to be less unorthodox, the pattern is substantially different, as can be observed when comparing Figure 4 to Figure F.

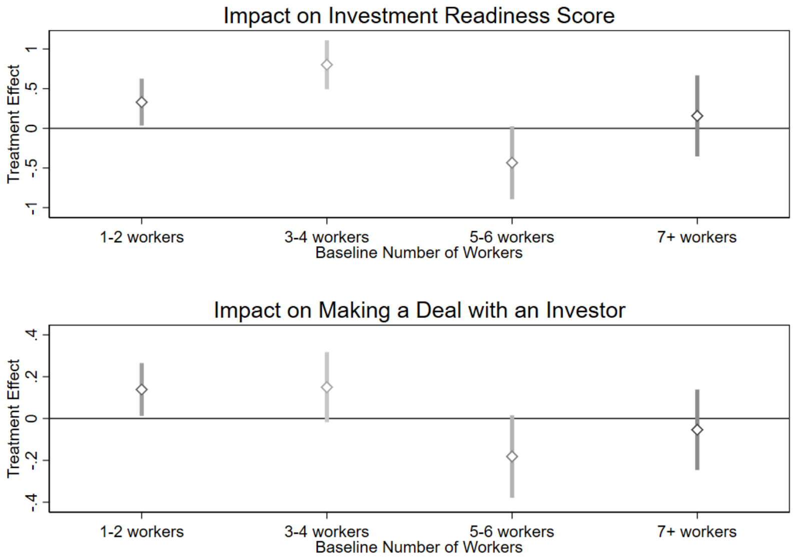

Figure 4.1: The Impact of the Program with Other Boundaries

Notes: The graphs show point estimates from rolling regressions which estimate the impact of being assigned to treatment. This figure shows the replication of Figure 2 in Cusolito et al. (2021) (Figure F in Appendix F), however, with another grouping of company sizes. This grouping suggests a change in the original conclusion that the IRP is less effective for larger firms to the IRP being less effective for 5-to-6-employee-sized firms. Please note that the different shades have no meaning.

In Figure 4, the main difference with Figure F lies in the increase of treatment effect on the investment readiness score (IRS) for firms larger than 7 employees. Consequently, the conclusion that the IRP is less effective for larger firms is suggested to be not robust to sensitivity analyses. The figure makes it seem that the sudden decrease for companies sized 5 to 6 is merely a dip. This dip is confirmed by Figure 4, where it is apparent that the effect of the treatment, participating in an IRP, only has a significantly different and negative effect on 6-employee firms. Furthermore, no clear trend can be observed in both figures. Hence, the theory is that the original conclusion is based on a dip caused by the 6-employee-sized firms, instead of the IRP actually having a lower effect on larger firms altogether. This motivates further research into the heterogeneity in the treatment effect of IRPs, using the ML methods used in Chernozhukov et al. (2023) and described in Section 2.

4.1 Heterogeneity Analysis

As mentioned, Algorithm 1 was implemented through Welz et al. (2022) in R for the analysis. Furthermore, in the implementation, fixed effects based on the strata used to split the data into treatment and control groups in Section 3 were included, and the standard errors were also clustered at the stratum level to account for intra-stratum correlation. This approach helps to control for unobserved heterogeneity within the strata. It ensures that the estimation accounts for the potential correlation of observations within the same strata, thereby providing more robust standard errors. Moreover, X_1 is controlled by proxy baseline estimates and proxy CATE estimates, and lasso regression is used as the learner of the propensity score. Also, $K = 5$ groups were created in the GATES and CATE analysis.

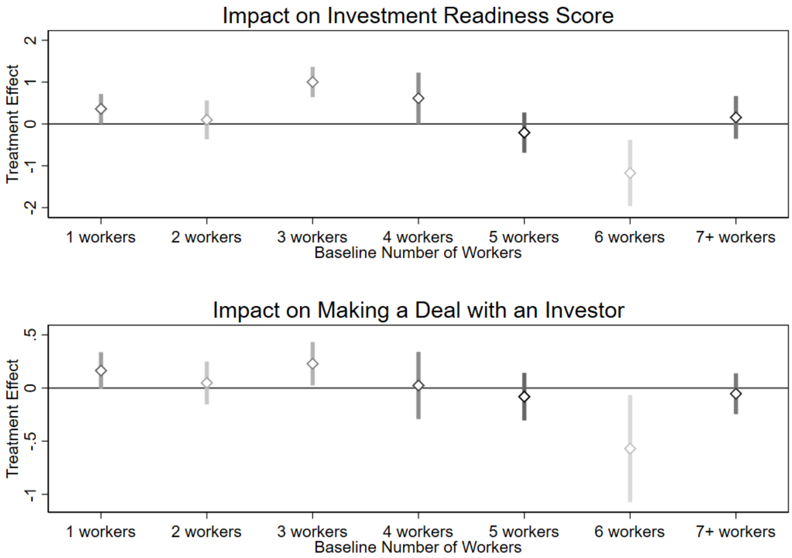

Figure 4.2: The Impact of the Program by Firm Size

Notes: The graphs show point estimates from rolling regressions which estimate the impact of being assigned to treatment. This figure shows the replication of Figure 2 in Cusolito et al. (2021) (Figure F in Appendix F), however, without grouping company sizes. This change suggests that the effect of participating in an IRP (the treatment) only has a significantly different effect on 6-employee firms when compared to other sizes. Please note that the different shades have no meaning.

In the analysis, the focus will mainly be on the treatment effect of two outcomes: the IRS and making a deal with an investor (MDI). In Table 4.1, five ML methods are compared for producing proxy predictors $S(Z_i)$ based on the highest Λ and $\bar{\Lambda}$ values. These methods were chosen due to their diverse strengths, providing a robust framework for understanding and predicting the treatment effects.

The first learning method is lasso regression, which makes use of variable selection and regularization. Hence, it is suitable for smaller datasets as it reduces overfitting. Secondly, random forests construct multiple decision trees and then combine their outputs, which is not only useful for high-dimensional data. For lower dimensions, it has the potential to perform well as it averages multiple trees to reduce variance. Thirdly, neural networks often perform extremely well for high-dimensional data to accurately tune the model. Yet, due to interest in its performance on smaller models, the method is included. K-means clustering partitions the data into distinct groups, which could reveal underlying patterns that may influence the treatment effect. For smaller models, it is more computationally efficient in comparison to other methods. Lastly, the linear model could perform reliably on smaller datasets due to its simplicity. In Cusolito et al. (2021), the heterogeneity analysis is completed using a linear regression model, hence including it allows fair comparison.

From Table 4.1, it can be concluded that the neural network outperforms in the GATES metric for both the IRS and MDI, providing the highest $\bar{\Lambda}$. For Λ, the best performing models differ, where the k-means method is best for IRS and lasso for MDI. Consequently, the focus will be on these three methods. Table 4.2 is the first instance where this focus continues, presenting the results of the BLP of CATE on the ML proxies. In the table, the estimates of the coefficients

β_1 and β_2, named average treatment effect (ATE) and heterogeneity loading (HET), respectively, are presented.

Table 4.1: Comparison of ML Methods: Investment Readiness Program

	Lasso	Random Forest	Neural Network	K-means	Linear Model
Investment Readiness Score					
Best BLP (Λ)	0.046	0.046	0.061	**0.064**	0.046
Best GATES ($\bar{\Lambda}$)	0.560	0.529	**0.877**	0.591	0.586
Making deal with investor					
Best BLP (Λ)	**0.023**	0.011	0.014	0.014	0.022
Best GATES ($\bar{\Lambda}$)	0.256	0.224	**0.277**	0.239	0.250

Notes: Medians over 100 splits. The Investment Readiness Score is the score given by judges in the semi-finals. The making a deal with an investor is based on the second survey, after two years of the program.

Table 4.2: BLP of Investment Readiness Programs Using Causal Proxies

	K-Means		Neural Network	
Investment Readiness Score	ATE	HET	ATE	HET
Estimate	0.024	0.359	0.058	0.021
90% Confidence interval	(-0.939, 0.895)	(-1.216, 1.840)	(-1.039, 1.135)	(-0.704, 0.656)
P-value	0.923	0.627	0.833	0.938
	Lasso		Neural Network	
Making deal with investor	ATE	HET	ATE	HET
Estimate	0.037	-0.168	-0.019	-0.013
90% Confidence interval	(-0.492, 0.555)	(-1.086, 0.858)	(-0.278, 0.361)	(-0.647, 0.558)
P-value	0.859	0.722	0.777	0.936

Notes: Medians over 100 splits. The Investment Readiness Score is the score given by judges in the semi-finals. The making a deal with an investor is based on the second survey, after two years of the program.

All estimates in Table 4.2 are near 0, suggesting there is no treatment effect and no heterogeneity. However, the values are insignificant meaning no conclusions can be made. Similarly, in Figure 4.1 the GATES of IRPs on IRS is presented. Even though all data points are within the ATE 90% confidence interval, the GATES values seem insignificant due to their wide confidence intervals. The same goes for Figure 4.1, where the values hover around the ATE estimates of approximately 0 with large intervals. Hence, no conclusion on heterogeneity can be supported by these tables and figures. For further interest, the figures are quantified in Table D.1 in Appendix D.

A conclusion can be based, however, on the widely ranged 90% confidence intervals. Namely, this range is indicative of a low signal-to-noise ratio in the data, or the methods are simply not able to properly "learn" the treatment effect heterogeneity. This is suspected to be due to the relatively small sample size, which leads to a power issue. Additionally, the variability in the treatment effects across subgroups may be complicated by noise, leading to the potential overfitting of these complex models to a small dataset. This can inflate the apparent variance of the estimates, hence not providing reliable insights. Another point to consider is the possibility of unobserved confounding variables, which highlights the importance of sensitivity analyses to assess their influence. However, these sensitivity analyses cannot be completed due to a lack of resources.

Furthermore, the first significant values are encountered in Table 4.3 and Table 4.4 reporting the CLAN estimates on the IRS and MDI, respectively. In the tables, only the p-values of the

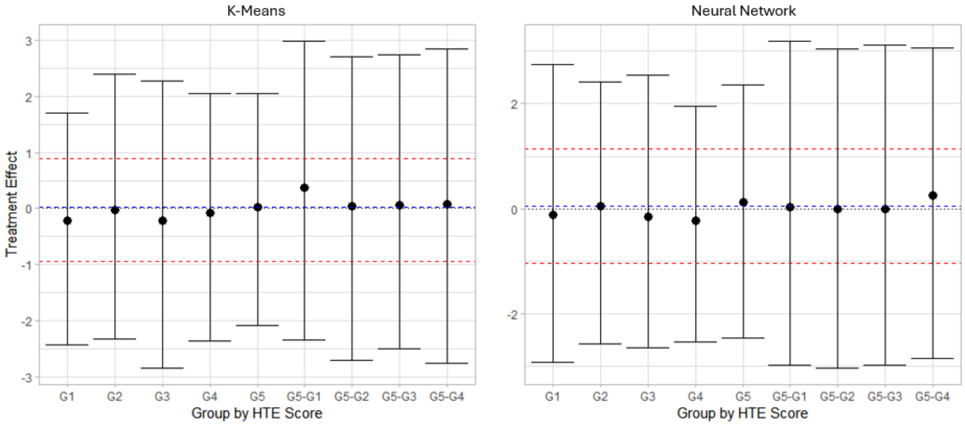

Figure 4.3: GATES of Investment Readiness Programs on Investment Readiness Score

Notes: GATES of IRPs on IRS, based upon the K-Means and Neural Network learners. In the figure, the blue and red dotted line(s) represent the ATE and confidence interval of the ATE, respectively. Similarly, the black dot and its confidence intervals represent the GATES. Moreover, the groups $G_1, \ldots, G_5$ each consist of 20% of the data, where G_1 and G_5 are the least and most affected groups, respectively. Furthermore, the variables that follow denote the difference between the most affected group and the other groups.

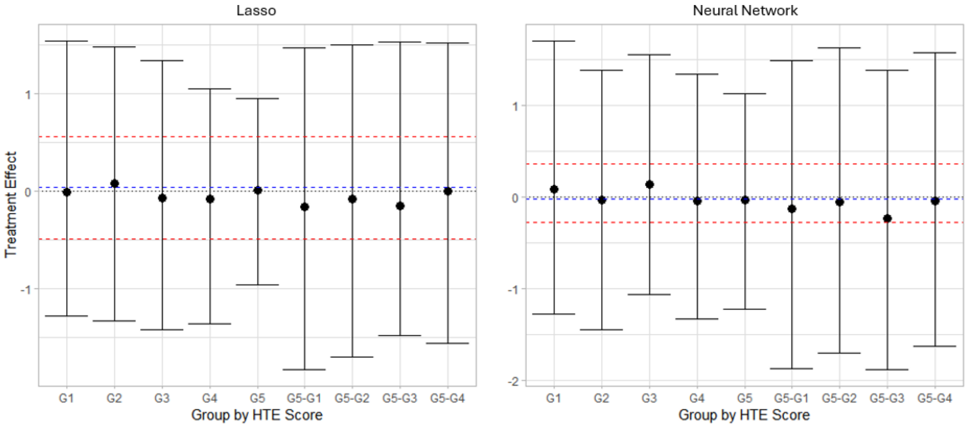

Figure 4.4: GATES of Investment Readiness Programs on Having Made a Deal with an Investor

Notes: GATES of IRPs on MDI, based upon the Lasso and Neural Network learners. In the figure, the blue and red dotted line(s) represent the ATE and confidence interval of the ATE, respectively. Similarly, the black dot and its confidence intervals represent the GATES. Moreover, the groups $G_1, \ldots, G_5$ each consist of 20% of the data, where G_1 and G_5 are the least and most affected groups, respectively. Furthermore, the variables that follow denote the difference between the most affected group and the other groups.

difference $\delta_5 - \delta_1$ are stated as the interest is in whether the most and least affected groups differ significantly. Logically, a significant difference would suggest the presence of heterogeneity. The selected covariates are chosen for presentation if they yield significant results, contribute to the investigation of heterogeneity among different company sizes, or offer noteworthy insights. To clarify, the new grouping presented in Figure 4 (in comparison to the figure from the original paper, given in Figure F in Appendix F) is used to ensure that there are adequate observations for each size.

Looking at Table 4.3, the difference in CLAN estimates of the k-means method for company sizes of 4 or more differ negatively (-0.320) and significantly (0.043). This verifies the conclusion by Cusolito et al. (2021), as the most affected group contains smaller firms significantly more

Table 4.3: CLAN of Investment Readiness Scores

	K-Means			Neural Network		
	20% Most	20% Least	Difference	20% Most	20% Least	Difference
	(δ_5)	(δ_1)	$(\delta_5 - \delta_1)$	(δ_5)	(δ_1)	$(\delta_5 - \delta_1)$
Investment Readiness Score						
Company Size: 4+	0.272	0.548	-0.320	0.341	0.548	-0.162
	(0.082, 0.463)	(0.305, 0.743)	(-0.603, -0.020)	(0.119, 0.517)	(0.305, 0.743)	(-0.502, 0.140)
			0.043			0.292
Company Size: 3-4	0.386	0.119	-0.221	0.388	0.167	0.262
	(0.158, 0.569)	(-0.011, 0.224)	(-0.036, 0.490)	(0.158, 0.569)	(0.000, 0.296)	(-0.014, 0.490)
			0.081			0.059
Company Size: 5-6	0.023	0.262	-0.241	0.068	0.214	-0.194
	(-0.032, 0.000)	(0.051, 0.425)	(-0.468, -0.051)	(-0.032, 0.135)	(0.018, 0.363)	(-0.425, 0.038)
			0.012			0.101
Company Size: 7+	0.114	0.167	-0.097	0.114	0.167	-0.066
	(-0.010, 0.214)	(0.000, 0.296)	(-0.305, 0.127)	(-0.010, 0.256)	(0.000, 0.296)	(-0.305, 0.135)
			0.401			0.564

Notes: Medians over 100 splits. Median Confidence Intervals ($\alpha = 0.05$) in parenthesis. P-values for the hypothesis over the parameter being equal to zero against the two-sided alternative under the intervals. The Investment Readiness Score is the score given by judges in the semi-finals.

than larger instances. Nonetheless, this analysis is not confirmed by the results from the neural network. Furthermore, with a quick look, the same can be said for company sizes 5 to 6, where the estimate is -0.241 with a p-value of 0.012. Hence, significantly more non 5-to-6-sized companies are in the most affected group, which verifies the theory of the instance being a dip. Again this is not verified by the estimates of the neural network. However, when looking closely at the interval of estimate δ_5 of the 5-to-6 employee firms of the k-means model, the max value does not cover the estimate itself, which is visualized in Figure D in Appendix D. This suggests a possible issue with the robustness or accuracy of the k-means model's estimates, likely indicating that the model's confidence intervals underestimate the variability in the data. Moreover, it might be possible the model is overfitting. Consequently, the issue raises concerns about the reliability and accuracy of the model as a whole.

As the k-means model seems to be not providing reliable estimates, the neural network is used to make conclusions. Its estimates are not significant for any covariate that was taken up in the model, which go beyond the four instances presented in the table. Hence, suggesting that there is no significant heterogeneity, even though it does tend towards the theory. As mentioned, neural networks need an extensive amount of data to recognize the intricate underlying patterns, which can explain why the estimates are not significant in this relatively small dataset. The limited sample size might have led to insufficient training, resulting in wide confidence intervals. Hence, the results in Table 4.3 suggest that the performance of the neural network is constrained.

Similar conclusions can be inferred from the results presented in Table 4.4, containing the CLAN estimates on MDI. Two significant results are notable. The first is that the difference for company size 5 to 6 is again negative (-0.382) and significant (0.001), and this time verified by the neural network model with an estimate of -0.242 and p-value of 0.048. Secondly, the estimate for the communication and collaboration industry is noted -0.287 with a p-value of 0.005. Even though the estimate (-0.242) in the neural network seems to support this, the p-value (0.282) determines it insignificant. Nonetheless, the confidence intervals for the covariates deem the lasso model to be unreliable, for similar reasons to discount the k-means model in Table 4.3. The intervals for covariates *company size: 5-6* and *communication & collaboration industry* can be visualized in Figure D and Figure D, respectively.

Once again, the results from the neural network are referred to for conclusions. The only significant estimate is that for firms of sizes 5 to 6, hence suggesting that there is significant heterogeneity for making a deal with an investor. As the result is negative, it suggests that the most affected group has significantly fewer 5-to-6-employee sized firms than the least affected group. Consequently, the theory is supported. Furthermore, the insignificant results imply no heterogeneity in receiving a high IRS by the judges. Either meaning that the judges have a flaw in their scoring, or that the firms do not live up to their expectations. Logically, the *judges investment readiness score* is based on a snapshot, meaning it does not have an insight on the entire framework of their judged company. Similar to Table 4.3, the confidence intervals are wide likely due to the small dataset, which might explain the insignificant results.

Table 4.4: CLAN of Making a Deal with an Investor

	Lasso			Neural Network		
	20% Most	20% Least	Difference	20% Most	20% Least	Difference
	(δ_5)	(δ_1)	($\delta_5 - \delta_1$)	(δ_5)	(δ_1)	($\delta_5 - \delta_1$)
Investor						
Company Size: 4+	0.386 (0.158, 0.569)	0.690 (0.460, 0.873)	-0.303 (-0.584, -0.012) 0.071	0.408 (0.158, 0.596)	0.548 (0.305, 0.743)	-0.206 (-0.502, 0.091) 0.173
Company Size: 3-4	0.386 (0.158, 0.569)	0.167 (0.000, 0.296)	0.167 (-0.110, 0.441) 0.206	0.391 (0.187, 0.595)	0.214 (0.019, 0.363)	0.126 (-0.164, 0.412) 0.369
Company Size: 5-6	0.023 (-0.010, 0.000)	0.405 (0.168, 0.594)	-0.382 (-0.618, -0.168) 0.001	0.079 (-0.024, 0.158)	0.310 (0.088, 0.505)	-0.242 (-0.484, 0.002) 0.048
Company Size: 7+	0.159 (0.000, 0.283)	0.119 (-0.033, 0.224)	-0.004 (-0.219, 0.206) 0.821	0.114 (-0.027, 0.214)	0.071 (-0.033, 0.141)	-0.003 (-0.185, 0.198) 0.974
Baseline Investment Readiness Score	2.818 (2.499, 3.168)	3.133 (2.820, 3.424)	-0.336 (-0.803, 0.127) 0.153	2.896 (2.559, 3.248)	3.043 (2.736, 3.366)	-0.100 (-0.598, 0.364) 0.687
Judges Investment Readiness Score	2.905 (2.546, 3.255)	3.242 (2.898, 3.580)	-0.375 (-0.884, 0.1267) 0.147	3.038 (2.678, 3.375)	3.183 (2.839, 3.547)	-0.141 (-0.659, 0.348) 0.561
Industry is Communication & Collaboration	0.000 (-0.010, 0.000)	0.262 (0.051, 0.425)	-0.287 (-0.513, -0.105) 0.005	0.068 (-0.032, 0.135)	0.179 (0.000, 0.296)	-0.127 (-0.339, 0.080) 0.282

Notes: Medians over 100 splits. Median Confidence Intervals ($\alpha = 0.05$) in parenthesis. P-values for the hypothesis over the parameter being equal to zero against the two-sided alternative under the intervals. The *Investor*variable refers to making a deal with an investor that is based on the second survey, after two years of the program.

Up to this point, one significant result has been obtained: the group affected most by participating in an IRP, in the sense of gaining an investor, contains significantly fewer firms with 5 to 6 employees when registering. As basing a conclusion on one result is insufficient and potentially unreliable, the heterogeneity effects in company sizes are further explored by comparing models on other various outcome metrics, as presented in Table 4.5. The estimates of $\delta_5 - \delta_1$ in the table are obtained by the neural network method, as it has been providing relatively more reliable and valid results. Despite the results still having wide 90% confidence intervals, two negative and significant estimates are noted. These are obtained for the external investment index and investment steps index. The result of the first index implies that the treatment has had significantly less effect on firms sized 5 to 6 in pushing them to take on new debt, make a deal with an outside investor, receive at least 25,000 euros in outside financing, or receive an incubator or accelerator grant. Hence, the already determined significant heterogeneity effect on *make a deal with an outside investor* is included. However, it is possible that this effect alone does not fully account for the significance of the index, as both the confidence interval (7.2% more narrow) and p-value (halved) have slightly improved. This suggests that the neural network may be capturing additional relevant patterns or that other outcomes within the index also contribute to its significance. Furthermore, the estimate of the investment steps index suggests that the treatment has inspired the 5-to-6-sized firms significantly less to contact, pitch to, or enter negotiations with outside investors, or have a mentor or expert support them to obtain financing. Furthermore, the other company size covariates might give insignificant estimates due to the small dataset, as explained several times above. Nonetheless, it can be stated with some degree of certainty that solely companies having 5 to 6 employees are less affected by the treatment.

Table 4.5: CLAN on Multiple Outcome Indices using Neural Networks

	Impact over Two Years						
	External Investment	Interested in Equity	Investment Steps	Specific Needs of Investors	General Investability	Firm Survival	Media Buzz Improvement
	$(\delta_5 - \delta_1)$	$(\delta_5 - \delta_1)$	$(\delta_5 - \delta_1)$	$(\delta_5 - \delta_1)$	$(\delta_5 - \delta_1)$	$(\delta_5 - \delta_1)$	$(\delta_5 - \delta_1)$
Company	-0.253	0.065	-0.122	-0.019	0.121	0.033	-0.071
Size: 4+	(-0.548, 0.034)	(-0.236, 0.371)	(-0.421, 0.166)	(-0.322, 0.282)	(-0.176, 0.421)	(-0.258, 0.330)	(-0.376, 0.233)
	0.090	0.665	0.423	0.909	0.296	0.673	0.646
Company	-0.287	-0.146	-0.238	-0.053	-0.148	-0.149	0.036
Size: 5-6	(-0.484 , -0.033)	(-0.339, 0.0565)	(-0.449, -0.018)	(-0.280, 0.172)	(-0.380, 0.057)	(-0.396, 0.075)	(-0.193, 0.248)
	0.024	0.197	0.048	0.658	0.197	0.161	0.772
Company	0.063	0.041	0.000	-0.008	0.179	0.110	-0.053
Size: 7+	(-0.175, 0.264)	(-0.175, 0.263)	(-0.186, 0.235)	(-0.219, 0.193)	(-0.052, 0.382)	(-0.095, 0.324)	(-0.275, 0.172)
	0.530	0.565	0.821	0.952	0.081	0.298	0.658

Notes: Medians over 100 splits. Median Confidence Intervals ($\alpha = 0.05$) in parenthesis. P-values for the hypothesis over the parameter being equal to zero against the two-sided alternative under the intervals. Please refer to Appendix F for the definitions of the indices.

This section will be concluded by arguing that the conclusion of IRPs having a lower effect on firms of 5 to 6 employees is solely due to firms with 6 employees. The original conclusion of Cusolito et al. (2021) stated that smaller firms had more benefit from the program, quantifying this in Table F.4. However, it was concluded with some certainty that only firms sized 5 to 6 have had a significantly lower treatment effect. Hence, narrowing the four or more employees range down to 5 and 6. Considering the only negative treatment effect in Figure 4, it strongly infers that the current focus can be further narrowed down to 6 employees. Nonetheless, this cannot be quantified precisely because further disaggregating the groups would make the sample sizes too small, further increasing the wide confidence intervals and likely resulting in no significant estimates. Hence, even though it is strongly suggested by the findings, the theory remains based on speculation.

5 Conclusion

In this paper, the impact of an IRP on access to equity finance for innovative start-ups and SMEs in the Western Balkans is addressed, and how the heterogeneous treatment effects vary across different firm characteristics, using machine learning and causal inference methods, as posed in the research question introduced earlier. A generic ML method is applied and evaluated to estimate and infer heterogeneous treatment effects to support the conclusions.

The original conclusion that IRPs are less effective for larger firms is not robust to sensitivity analysis, spurring the theory that this is only the case for firms with 6 employees, which can be supported using the results from the neural network model. Significant heterogeneity in the treatment effect for company sizes is found. From the results, it is concluded with certain robustness and reliability that only firms with 5 to 6 employees are less affected by participating in an IRP. This group cannot be disaggregated further as it makes the sample size too small. However, it can be strongly argued with visual evidence that this result is due to 6 employee firms. Furthermore, the insignificant results might be due to the small dataset, as suggested by the wide confidence intervals in both the balance tests and estimations. This can also be inferred from the k-means and lasso models producing seemingly unreliable and inaccurate results, as is implied by confidence intervals that do not include the estimation.

Moreover, it is notable that none of the results of the neural network support a significantly different effect for company sizes of more than 4. This insignificance is surprising considering that the original paper did find such significant results using simple linear regression models with fixed effects. However, the linear model in these results did not perform well. A reason for this is that the best-performing models were based on the highest Λ (BLP) and $\bar{\Lambda}$ (GATES) values. While GATES does measure the ability to detect heterogeneity to some extent by comparing treatment effects across subgroups, it is not a broad measure of a model's capacity to uncover all forms of heterogeneity in the data. The effectiveness of GATES in detecting heterogeneity depends on how well the subgroups are defined and how accurately the model captures these subgroups. Consequently, the CLAN estimates may not fully reflect the differences in treatment effects across various firm sizes. This could also be responsible for the broader confidence intervals and a lack of significant findings. Hence, it can be concluded that the generic ML method is not effective for small datasets, as more resources were needed for a similar result when using simple linear regression.

A possible reason for the significantly lower treatment effect for sizes 5 to 6 might be that start-ups in this size category tend to expand too quickly, likely resulting in overextension. Hence, they may need to scale back to stabilize again, which slows down the progress being made. Another explanation might have been that this group size has already made significantly more deals with outside investors at the point of registering, resulting in not wanting to give up even more equity. However, by examining the data it is observed that this is not the case.

To further answer the research question, it seems companies sized 5 to 6 have no benefit of participating in the IRP over the online investment readiness course. In general, the treatment effect of an IRP is not significantly higher than an online course, suggesting an insignificantly different impact on access to equity finance for innovative start-ups and SMEs in the Western Balkans. This could only be strengthened when considering the cost difference, which should be further analyzed.

Therefore, for future research, a cost-effectiveness analysis is suggested that will likely further diminish the effectiveness of an IRP compared to the online course. Furthermore, to properly apply the generic ML method, it is necessary to extend the dataset by having more firms participate. This extension is challenging due to financing and a limiting number of start-ups and SMEs present in the Western Balkans. Lastly, it might be interesting to broaden the number of years the companies are tracked. Again, this could be challenging due to financing issues.

References

Athey, S. & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353-7360. Retrieved from `https://www.pnas.org/doi/abs/10.1073/pnas.1510489113` doi: 10.1073/pnas.1510489113

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2017). *Double/debiased machine learning for treatment and causal parameters.*

Chernozhukov, V., Demirer, M., Duflo, E. & Fernández-Val, I. (2023). *Fisher-schultz lecture: Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in india.*

Cusolito, A. P., Dautovic, E. & McKenzie, D. (2021, 07). Can Government Intervention Make Firms More Investment Ready? A Randomized Experiment in the Western Balkans. *The Review of Economics and Statistics*, *103*(3), 428-442. Retrieved from `https://doi.org/10.1162/rest_a_00882` doi: 10.1162/rest_a_00882

Del Sarto, N., Cruz Cazares, C. & Di Minin, A. (2022). Startup accelerators as an open environment: The impact on startups' innovative performance. *Technovation*, *113*, 102425. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0166497221002066` doi: https://doi.org/10.1016/j.technovation.2021.102425

Donders, A. R. T., van der Heijden, G. J., Stijnen, T. & Moons, K. G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087-1091. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0895435606001971` doi: https://doi.org/10.1016/j.jclinepi.2006.01.014

Gattini, L., Soljan, D., Hauser, P., Dolgaya, T., Revenko, S. & Kortenbusch, M. (2016). *Assessment of financing needs of smes in the western balkans countries. synthesis report.* European Investment Bank.

Genovese, C. R. & Wasserman, L. (2007). *Adaptive confidence bands.*

Levine, R. J. (1981). *Ethics and regulation of clinical research.* Baltimore: Urban & Schwarzenberg.

Liang, T., Rakhlin, A. & Sridharan, K. (2015, 03–06 Jul). Learning with square loss: Localization through offset rademacher complexity. In P. Grünwald, E. Hazan & S. Kale (Eds.), *Proceedings of the 28th conference on learning theory* (Vol. 40, pp. 1260–1285). Paris, France: PMLR. Retrieved from `https://proceedings.mlr.press/v40/Liang15.html`

Mallaby, S. (2023). *The power law: Venture capital and the art of disruption.* London, UK: Penguin Books Ltd (UK). (Refer to page 401)

Mason, C. & Kwok, J. (2010). Investment readiness programmes and access to finance: A critical review of design issues. *Local Economy*, *25*(4), 269-292. Retrieved from `https://doi.org/10.1080/02690942.2010.504570` doi: 10.1080/02690942.2010.504570

Vizjak, A. & Vizjak, M. (2016, june). Transatlantic Trade And Investment Partnership (Ttip) And Regulation Of Financial Markets. *Economic Thought and Practice*, *25*(1), 319-336. Retrieved from https://ideas.repec.org/a/avo/emipdu/v25y2016i1p319-336.html

WBIF. (2024). *Wb edif - western balkans enterprise development & innovation facility.* Retrieved from https://www.wbif.eu/wb-edif (Accessed: 2024-06-24)

Welz, M., Alfons, A., Demirer, M. & Chernozhukov, V. (2022). Genericml: Generic machine learning inference [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=GenericML (R package version 0.2.3)

World Bank. (2024). *Research and development expenditure (% of gdp) - croatia.* Retrieved from https://data.worldbank.org/indicator/GB.XPD.RSDV.GD.ZS?locations=HR (Accessed: 2024-06-24)

A Methodology Extension

A.1 Emperical Estimation of the BLP Model

To empirically estimate the Equation 3:

$$Y_iH_i = \hat{\mu}_0' X_{1i}H_i + \hat{\mu}_1 + \hat{\mu}_2(S_i - \mathbb{E}_{N,M}[S_i]) + \hat{\epsilon}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[\hat{\epsilon}_i \tilde{X}_i] = 0, \tag{10}$$

where $\tilde{X}_i := (X_{1i}', \tilde{X}_{2i})'$, $\tilde{X}_{2i} := (1, S_A(Z)_i - \mathbb{E}_{N,M}\, S_A(Z)_i)'$, and $X_{1i} := [1, B(Z_i), p(Z_i), p(Z_i)S_A(Z_i)]'$. To remind and clarify from the introductory part of the methodology, N denotes the total number of observations and M is the main sample. Hence, the operator $\mathbb{E}_{N,M}$ refers to the empirical expectation concerning the main sample.

A.2 Emperical Estimation of the GATES Model

To empirically estimate the Equation 4, the following is used

$$Y_iH_i = \hat{\mu}_0 X_{1i}H_i + \hat{\mu}\tilde{W}_i + \hat{\nu}_i, \quad i \in M, \quad \mathbb{E}_{N,M}[\hat{\nu}_i \tilde{W}_i] = 0, \tag{11}$$

where $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_K)'$. To remind and clarify from the introductory part of the methodology, N denotes the total number of observations and M is the main sample. Hence, the operator $\mathbb{E}_{N,M}$ refers to the empirical expectation concerning the main sample.

A.3 Model Selection Extension

In this subsection of the appendix, various methods are introduced for determining the better model, when multiple methods perform similarly according to Λ and $\bar{\Lambda}$.

For instance, a combined confidence set can be constructed to maintain inferential coverage. Another possibility is to use a Bonferroni adjustment to modify the confidence level if the selection is based on empirical results. Even though this is not the case in this paper, for large datasets, further data splitting can also help determine the best-performing ML method before applying the proxies to the main sample. Despite that, this paper does not make use of such confidence sets or other techniques to distinguish between the two top performers.

A.4 Inference Methods

In this subsection of the appendix, robust post-processing methods are introduced, focusing mostly on inferences that account for the uncertainty arising from parameter estimation and sample splitting. These methods are crucial for ensuring the validity and reliability of the results. Even though this subsection is less extensive than Section 2.1, it introduces several parameters and variables. First, let θ denote a generic target parameter that can take forms from $\theta = \beta_2$ as the heterogeneity loading parameter to $\theta = \delta_K - \delta_1$ as the difference in the expectation of the characteristics of the most and least impacted groups in CLAN. Furthermore, let (a, m) denote a fixed partition of $1, \ldots, N$. Again, probabilities, expectations, estimands, and ML proxies can be conditional on the auxiliary sample a, $Data_a := \{(Y_i, D_i, X_i)\}_{i \in a}$. However, Chernozhukov et al. (2023) makes clear that "this dependence vanishes as the size of the auxiliary sample becomes large under suitable conditions".

A.4.1 From Single Split to Multiple Splits

Starting with a single split of data induced by the partition $\{(a, m)\}$ of $1, ..., N$ into sets of cardinality $(N - n, n)$. Each example admits an estimator $\hat{\theta}_a$ that is approximately Gaussian, given $Data_a$, as $(N - n, n) \to \infty$ and for any z,

$$\mathrm{P}(\hat{\sigma}_a^{-1}(\hat{\theta}_a - \theta_a) < z \mid \mathrm{Data}_a) \to_P \Phi(z). \tag{12}$$

Consequently, the standard p-values for testing the null hypothesis $\theta_a = \theta_0$ against the alternatives $\theta_a > \theta_0$ and $\theta_a < \theta_0$ are nearly uniform under the null. Furthermore, the standard confidence interval (CI) covers θ_a with an approximate probability of $1 - \alpha$ given $Data_a$. Hence, the inference is straightforward on a single data split.

With a single data split various factors of uncertainty, such as variability in the estimators, potential biases, and the randomness of the split itself, come into play. To reduce these factors, multiple splits (a, m)'s are used. In Chernozhukov et al. (2023), the authors propose quintile aggregation methods and analyze their properties to aggregate the results across the collection of splits, which is defined in Definition 5.

Definition 5 (Collection of Splits). Consider the collection $\{(a, m), a \in \mathcal{A}\}$ of partitions of $[N] = \{1, \ldots, N\}$ into auxiliary sets a of cardinality $N - n$ and main sets m of cardinality n. The collection is generated independently of

$$Data := (Y_i, D_i, X_i)_{i=1}^N.$$

Let estimand θ_A be a random variable conditional on $Data$, where A is a uniform distribution variable taking values $a \in \mathcal{A}$. To clarify, $A\ U(\mathcal{A})$. As θ_a is merely the estimand for a specific partition, this paper rather focuses its inference on the median value of θ_A. This preference lies in θ_A being associated with a random partition, which better represents the variability across all possible partitions. Logically, each partition yields a different estimator $\hat{\theta_A}$ and their corresponding distributions. Consequently, given $Data$, these estimators $\hat{\theta_A}$, but also p-values p_A, and intervals $[L_A, U_A]$ are all random variables. To aggregate these results, quantile aggregation methods are used.

To delve deeper into the above, the median-aggregated p-value is defined in Definition 6, followed by the quantile-aggregated point and interval estimators in Definition 7.

Definition 6 (Median-Aggregated P-value). The median p-values for testing one-sided alternative hypotheses are

$$p^{\pm} = M(p_A^{\pm} \mid Data).$$

The two-sided median p-value is $\bar{p} = 2\min(p^+, p^-)$.

In this context, $p_A^{\pm}$ represents the p-value from the A-th split for testing the one-sided alternative hypothesis. The function $M(\cdot \mid Data)$ denotes the central median computed over the distribution of p-values given the observed data.

Definition 7 (Quantile-Aggregated Point and Interval Estimators). The median point estimator is:

$$\hat{\theta} := \mathrm{M}(\hat{\theta}_A \mid Data).$$

The β-quantile confidence interval is $[L, U]$, where

$$L := Q_\beta(L_A \mid Data), \quad U := Q_{1-\beta}(U_A \mid Data), \quad \beta \leq 1/2.$$

These definitions can be understood as providing inferential summaries that reduce risk.

A.4.2 Conditions for the Understanding and Interpretation of Two Key Inferential Theorems

In this subsection, three regularity conditions are discussed that are essential for understanding two inferential theorems that are also debated. The first theorem concerns the *uniform validity of median-aggregated p-value* and the second covers the *properties of the confidence interval for* θ^*. Subsequently, the aforementioned conditions, labeled as (R1), (R2), and (R3), establish the necessary regularity and concentration properties for the inferential results and theorems to hold.

In Chernozhukov et al. (2023), the first condition (R1) is known as the main or *principal regularity condition*, which assumes approximate normality of the split-sample t-statistics:

(R1) There exist a sequence of positive constants $\gamma'_N \searrow 0$ as $(n, N-n) \to \infty$, such that

$$\sup_{z \in R} \left| \mathrm{P}\left\{ \hat{\sigma}_A^{-1}(\hat{\theta}_A - \theta_A) < z \right\} - \Phi(z) \right| \leq \gamma'_N. \tag{13}$$

As can be observed, this condition ensures that the distribution of the standardized estimator $\hat{\sigma}_A^{-1}(\hat{\theta}_A - \theta_A)$ converges uniformly to the standard normal distribution as the sample sizes n and $N-n$ grow large. Hence, the condition implies that the estimation errors, when standardized, become normally distributed, which is useful as it allows the use of normal distribution properties for inference.

Subsequently, in the context of hypothesis testing with multiple splits, the properties of median p-values are evaluated to ensure robustness against various alternatives. Condition (R2) is beneficial with this evaluation as it provides a framework for assessing the accuracy of these median p-values. The condition particularly focuses on their behavior within specified nominal levels. In (R2) the properties of the median *p*-values under (R1) are established, where a concentration condition is used for approximate medians.

(R2) For all $z = \Phi^{-1}(\alpha)$, where the nominal level of interest α is in some closed sub-interval of $(0, 1/4)$, and some sequences of positive constants $\gamma''_N \searrow 0$ and non-negative constants $\varepsilon_N \searrow 0$:

$$\begin{aligned}
\mathrm{P}\left(Q_{.5-\varepsilon_N}\left(\hat{\sigma}_A^{-1}(\theta_A - \hat{\theta}_A) \mid Data \right) < z \right) &\leq \mathrm{P}\left(\hat{\sigma}_A^{-1}(\theta_A - \hat{\theta}_A) < z \right) + \gamma''_N, \\
\mathrm{P}\left(Q_{.5-\varepsilon_N}\left(\hat{\sigma}_A^{-1}(\hat{\theta}_A - \theta_A) \mid Data \right) < z \right) &\leq \mathrm{P}\left(\hat{\sigma}_A^{-1}(\hat{\theta}_A - \theta_A) < z \right) + \gamma''_N.
\end{aligned} \tag{14}$$

The condition makes it apparent that the approximate median of the t-statistics across multiple splits tends to be more concentrated than the t-statistic from any individual split, implying a stronger form of convergence. Intuitively, this means that using the median of the split-specific estimates leads to better inference properties compared to using any individual split.

Theorem 4.1 (Theorem 3 in Appendix B.3) demonstrates that under the median concentration condition, the median p-values have the standard property. Without this condition, the p-values need to be multiplied by 2 to account for the potential overestimation of the test statistics' significance. Hence, ensuring robustness against false positives.

Before moving to the final condition, the inferential target is introduced in Definition 8.

Definition 8 (Inferential Target). Our inferential target is the median estimand:

$$\theta^* = \mathrm{M}(\theta_A \mid Data).$$

The target is appealing because it represents a typical value, and in many cases, θ_A will cluster around its median θ^*, reinforcing its suitability. The following and last condition (R3) covers these many cases of clustering.

(R3) For some positive sequences of constants $r_N \searrow 0$ and $\gamma'''_N \searrow 0$ as $(n, N - n) \to \infty$, the following concentration condition

$$\mathrm{P}(\hat{\sigma}_A^{-1} |\theta^* - \theta_A| > r_N) \leq \gamma'''_N. \tag{15}$$

This last condition is according to Chernozhukov et al. (2023) "high-level". Namely, the condition ensures that the estimator θ_A concentrates around the target parameter θ^* at a rate faster than the estimation error $\hat{\sigma}_A$ as the sample size grows. Hence, it implies that the probability of large deviations of the estimator from the target parameter decreases. This in turn enhances the reliability of the inference as the sample size increases.

With the three conditions in place, the properties of the confidence interval are established for θ^*, as is presented in Theorem 4.3 (Theorem 4 in Appendix B.4). From the theorem, it can be observed that with the strongest assumptions, the probability that the target θ^* is covered is at least $1 - \alpha - o(1)$. With minimal assumptions, the coverage probability is $1 - 2\alpha - o(1)$. Numerical results in Chernozhukov et al. (2023) demonstrated that the confidence intervals tend to be conservative, often providing coverage that exceeds $1 - \alpha$. Consequently, based on this evidence, the authors recommend using $1 - \alpha$ as the nominal level.

A.5 Oracle Properties

In this subsection of the appendix, the theoretical properties that arise from the approach of the causal learner defined in Equation 9 are further detailed. As mentioned in Section 2.3, the properties show that causal learners can closely approximate the true CATE function without direct observation.

Equation 9 enhances standard predictive learners by providing better approximations for $E[Y|D, Z]$ and help validate and select machine learning methods for targeting the CATE function. It refines the strategy of Athey & Imbens (2016) by introducing denoising, making it more robust without relying on a consistent estimation of the regression function, unlike related HT strategies.

Theorem 5.1 (Theorem 5 in Appendix B.5) illustrates that the minimizers of the two loss functions yield the best approximation, in the mean-square sense, to the actual CATE function $s_0(Z)$ within the class $\mathcal{S}$. Despite not observing $s_0(Z)$ directly, this performance is often referred to as an "oracle." For these minimizers to exist, it is sufficient if the set $\mathcal{S}$ is convex and closed in the $L^2(P)$ norm.

In Chernozhukov et al. (2023), the authors visually compare predictive and causal CATE learners solving learning objectives in Equation 9 for Random Forests (RF), Neural Networks (NN), OLS, and Causal Forest (CS). From their figures, it is easily observed that causal learners are better at approximating the CATE function for all models. Hence, providing better proxies for CATE in the process. Consequently, the authors concluded that the causal learners improved the RMSE for RF and NN by a range from 21% to 44%, for OLS about 35%, and for CF it ranges between 54% to 63%.

The remainder of this subsection will cover learning guarantees of the causal learner in Equation 9. These guarantees are derived from advanced statistical learning theory Liang et al. (2015), particularly through the application of the expected offset Rademacher complexity (ORC). This ORC quantifies the complexity of the functional class by measuring its ability to fit i.i.d. Rademacher noise, which consists of random variables that independently take values -1 and 1 with probability 0.5. A higher ORC indicates a more complex function class.

According to Liang et al. (2015), the expected ORC of the function class $\mathcal{H}$ is defined as:

$$\mathcal{R}^o(\mathcal{A}, \mathcal{H}, c) := E \sup_{h \in \mathcal{H}} \frac{1}{|A|} \sum_{i \in A} [e_i h(Z_i) - c h(Z_i)^2],$$

where $\{e_i\}$ are i.i.d Rademacher noise variables and $c > 0$ is a positive constant. Furthermore, Liang et al. (2015) further establishes that ORC provides the sharpest characterization of complexity, showing it is bounded by the critical radii of function classes defined through local Rademacher complexity or standard uniform covering entropy. These measure the ability of a function class to fit random noise and its overall size, respectively. The bounding further implies that ORC provides a precise measure of complexity. In ML, this tends to be useful in selecting and tuning models, making ORC useful in modern ML.

The last theorem, Theorem 5.2 (Theorem 6 in Appendix B.6), establishes formal learning guarantees for causal learners where the assumption of s_0 being consistent is not present. The result the theorem suggests that a smaller ORC for the functional parameter spaces corresponds to a lower excess risk of the estimator compared to the oracle approximation of the CATE. Furthermore, in Chernozhukov et al. (2023) the authors also conclude that "learning the technical baseline function does not affect the rate of learning the oracle prediction $s\bullet$"; and the ORC is determined solely by the complexity of $\mathcal{S}$. Lastly, loss function 9 can be used for choosing the best ML method.

B Theorems

B.1 BLP Identification

Using Theorem 1, the linear projection in 3 identifies the BLP, and the theorem allows me to estimate the equation empirically.

Theorem 1 (BLP Identification). Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y has finite second moments, $\tilde{X}$ is such that $E_A[\tilde{X}\tilde{X}']$ is finite and full rank, which requires $\mathrm{Var}(S_A(Z)) > 0$. Then, (μ_1, μ_2) defined in Equation 3 identifies the coefficients of the BLP,

$$\mu_1 = \beta_1, \quad \mu_2 = \beta_2.$$

B.2 GATES Identification

Theorem 2 (GATES). Consider $z \mapsto S(z)$ and $z \mapsto B(z)$ as fixed maps. Assume that Y has finite second moments and W and $\tilde{W}$ are such that $E_A[WW']$ and $E_A[\tilde{W}\tilde{W}']$ are finite and have full rank. Consider $\alpha = (\alpha_k)_{k=1}^K$ defined by the weighted regression equation from *Strategy A*

$$Y = \alpha_0' X_1 + \sum_{k=1}^{K} \alpha_k \cdot [D - p(Z)] \cdot 1(G_k) + \nu, \quad E_A[w(Z)\nu W] = 0$$

and $\mu = (\mu_k)_{k=1}^K$ defined by the regression Equation 4. These coefficients are equal and identify the GATES:

$$\alpha_k = \mu_k = \gamma_k = E_A[s_0(Z)|G_k], \quad k = 1, \ldots, K.$$

B.3 Uniform Validity of Median-Aggregated P-Value

Theorem 3 (Uniform Validity of Median-Aggregated P-Value). Suppose that the null hypothesis $H_0 : \theta_A = \theta_0$, which implies that θ_A does not vary with a, holds with probability one. Let p be either of $\{p^+, p^-, \bar{p}\}$. (i) Suppose that approximate normality (R1) holds, then

$$\mathrm{P}(2p < \alpha) \leq \alpha + o(1),$$

where the $o(1)$ depends only on γ_N'. (ii) Suppose in addition that the median concentration condition (R2) holds with $\varepsilon_N = 0$, then

$$\mathrm{P}(p < \alpha) \leq \alpha + o(1),$$

where the $o(1)$ depends only on γ_N' and γ_N''.

B.4 Properties of the Confidence Interval for θ^*

Theorem 4 (Properties of the Confidence Interval for θ^*). Let $\beta = 1/2$. (i) Suppose that (R1) and (R3) hold. Then,

$$\mathrm{P}(\theta^* \in [L, U]) \geq 1 - 2\alpha - o(1),$$

where $o(1)$ depends only on γ'_N, γ'''_N and r_N. (ii) Suppose in addition that (R2) holds with $\varepsilon_N = 2\sqrt{\gamma'''_N}$. Then,

$$\mathrm{P}(\theta^* \in [L, U]) \geq 1 - \alpha - o(1),$$

where $o(1)$ depends only on $\gamma_N, \gamma''_N, \gamma'''_N, r_N$. (iii) In either case, the event $\theta^* \in [L, U]$ implies $|\theta^* - \hat{\theta}| \leq |U - L|$.

B.5 Oracle Properties of the Population Objective Function

Theorem 5 (Oracle Properties of the Population Objective Function). Suppose that Y, $b(Z)$, $s(Z)$, and $w(Z)$ are square integrable. (1) Then, the expectations of the loss functions in 9 are

$$E[YH - b(Z)H - s(Z)H]^2 = E[s_0(Z) - s(Z)]^2 + C_b, \tag{16}$$

where $C_b := E[w(Z)(\tilde{b}_0(Z) - b(Z))^2] + C$ for some constant C. (2) Therefore, the minimizer, say $s_\bullet(Z)$, of the left-hand side of Equation 16 over $s \in \mathcal{S}$, if it exists, also minimizes the oracle loss function $E[s_0(Z) - s(Z)]^2$ over the same set.

B.6 Near-Oracle Guarantees for Causal Learners

Theorem 6 (Near-Oracle Guarantees for Causal Learners). Suppose that Y, the elements of $\mathcal{B}$ and $\mathcal{S}$, and $w(Z)$ are bounded in absolute values by K, and $\mathcal{B}$ and $\mathcal{S}$ are closed, convex, and symmetric sets. The estimator S obtained as a solution of either (A) or (B) is as good as using the best in class approximation, say $s_\bullet(Z)$, up to an error expressed in terms of ORC:

$$0 \leq E[s_\bullet(Z) - S(Z)]^2 \leq E[s_0(Z) - S(Z)]^2 - E[s_0(Z) - s_\bullet(Z)]^2 \leq C_K \mathcal{R}^o(\mathcal{A}, \mathcal{H}, c_K), \tag{17}$$

where C_K and c_K are positive constants that only depend on K, $\mathcal{H} := 4(w(Z)^2\mathcal{B} + Hw(Z)\mathcal{S})$ for type (A) loss, and $\mathcal{H} := 4(H\mathcal{B} + w(Z)\mathcal{S})$ for type (B) loss.

C Dealing with NA Values

This section in the appendix explains the imputation methods used in this paper, contains the figures used to classify the data MCAR, MAR, or MNAR, and which imputation method performs best, as is shortly discussed in Section 3.

Please note that the variable names in the figures that contain missing values are defined by Cusolito et al. (2021) as follows. First of all, the variables containing only *f1* are from the first survey, whereas *fu2* or *fu2_f1* refer to the second follow-up survey. **Media Buzz** (*b_mediabuzz2016*, *b_mediabuzz2017*) is a standardized index of whether the firm is mentioned in the media, the number of media mentions, number of Facebook likes and number of Twitter followers in the respective year (2016 or 2017). **Firm survival** (*f1_operate*, *fu2_operate*) is a binary variable that takes value one if the firm is operating, and zero otherwise. **Interested in equity** (*f1_interestindex*, *fu2_f1_interestindex*) is a standardized index of whether the firm is interested in equity financing, the maximum equity share they are willing to have owned by outside investors, whether they have specific deal terms for investors, and whether they would consider a royalty-based investment. **General investability** (*f1_generalindex*, *fu2_f1_generalindex*) is a standardized index of number of employees, whether the founders work full-time in the business, whether the firm had positive sales in the first quarter of the year, whether total sales exceed 10,000 euros in that quarter, whether the firm made a positive profit in the past year, and whether the firm made sales to Western Europe or the United States. **Specific needs of investors** (*f1_specificindex*, *fu2_f1_specificindex*) is a standardized index of whether business and personal accounts are separated, whether the firm has made a revenue projection for the next year, whether it knows customer acquisition costs, the number of key metrics tracked, whether it has found out if the product or service can be covered by intellectual property protection, and whether it has at least one form of intellectual property protection received or pending. **Investment steps** (*f1_stepindex*, *fu2_f1_stepindex*) is a standardized index of having contacted an outside investor, made a pitch to an outside investor, have a mentor or external expert supporting them to obtain financing, and entered into negotiations with an outside investor. **External investment** (*f1_externalindex*, *fu2_f1_externalindex*) is a standardized index of having taken on new debt, having made a deal with an outside investor, have received at least 25,000 euros in outside financing, and have received an incubator or accelerator grant (all since August 2015). **Made a deal with an investor** (*f1_external2*, *fu2_f1_external2*) indicates having made a deal with an outside investor since August 2015 (program start).

C.1 Imputation

To fill the unanswered survey questions, imputation is used as it provides unbiased estimates of missing values by predicting them based on available data (Donders et al., 2006). These substituted values can be derived from similar observations, predictive models, or statistical techniques. Before being able to use such methods, it is important to understand the mechanism behind the missingness, where it is key that the data contains no multicollinearity. Knowing this mechanism guides the choice of the imputation method, hence ensuring the validity of the statistical inferences drawn from the data. The mechanisms can be classified into three categories, according to Donders et al. (2006). Donders and his team state that the missing data is classified as missing completely at random (MCAR) when “subjects who have missing data are a random subset of the complete sample of subjects.” Furthermore, the authors stamp the data as missing not at random (MNAR) when “the probability that an observation is missing depends on information that is not observed.” However, the authors state that most cases of missing data are neither of the above, but missing at random (MAR) instead. This classification refers to the reason for missingness being based on other observation characteristics.

The pattern of the data used in this paper follows the common way, being classified as MAR. This conclusion is taken from the figures that are placed in Appendix C.2, which are placed there as it is not the focus of the paper. As a result, the data is imputed using k-nearest neighbors (kNN) and multiple imputation by chained equations (MICE). The former replaces a missing value with the most frequent value among the k-nearest neighbors, the latter generates multiple complete datasets by imputing missing values iteratively. In comparison, the kNN imputation method performed better as it provides more predictable results, with the root mean squared error (RMSE) of kNN being 0.397 compared to the MICE RMSE of 0.413 of a simple LM regression. Even though the difference is not significant, the kNN density plot is more in line with the original sample, based on Figure C.3 in Appendix C.3. Hence, the kNN imputed data is solely used.

C.2 MCAR, MAR, or MNAR

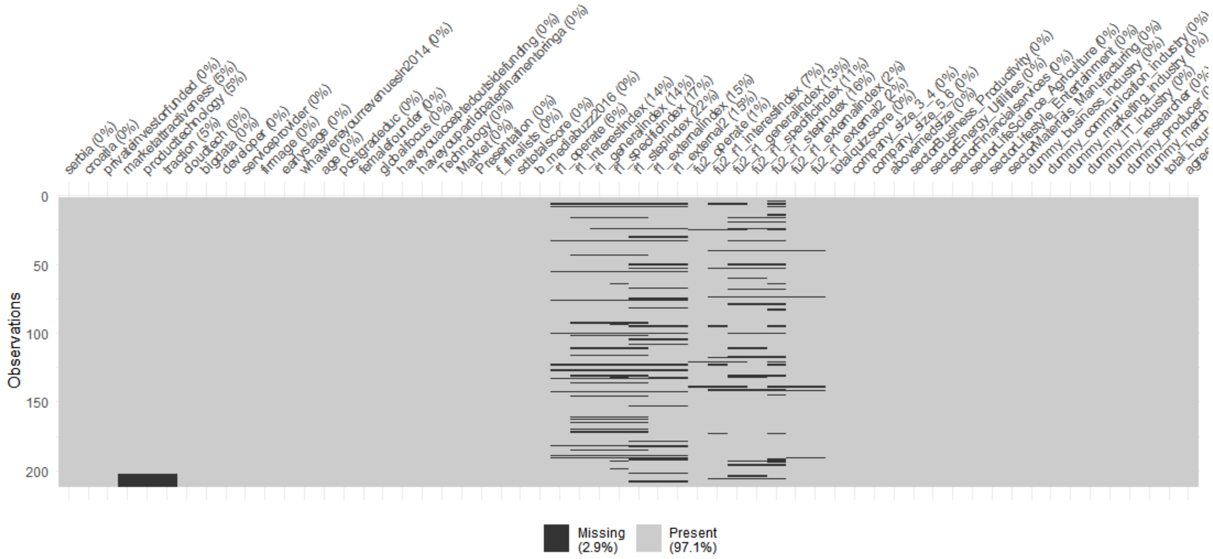

Figure C.1: Vizualization of Observations Missing Data

Notes: The missing parts of an observation are denoted black, which is only 2.9%. Generally, these missing values are in the same region, however, they are scattered in a seemingly random order across this area. For an explanation of the relevant variables, please refer to the introduction of Appendix C.

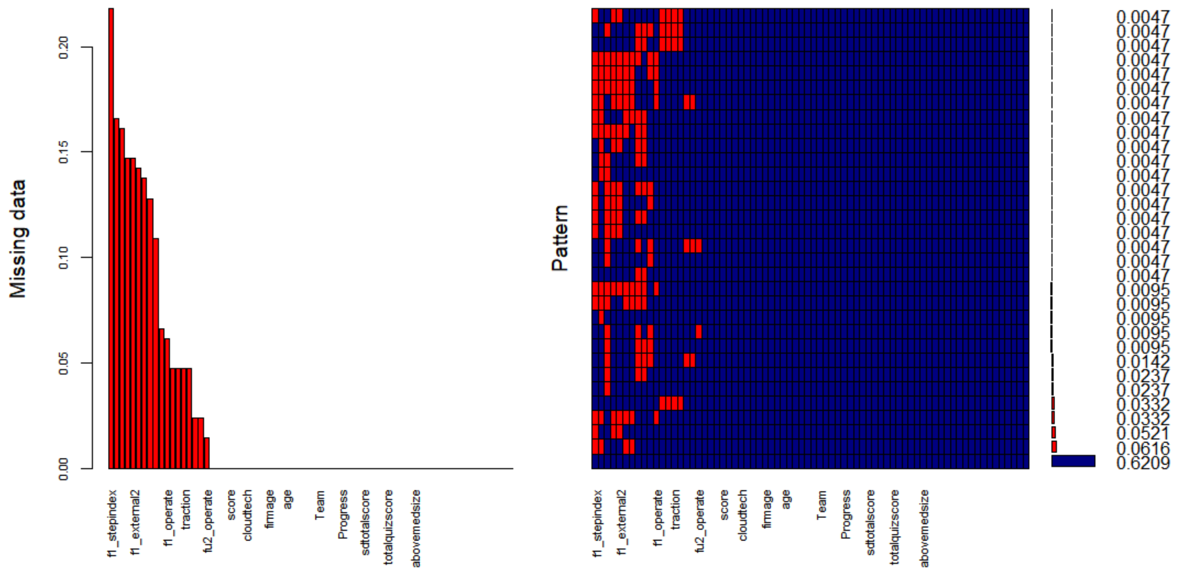

Figure C.2: Missing Data Heatmap

Notes: The left panel is a bar chart showing the proportion of missing data for each variable. The right panel is a missing data pattern plot, where each row represents an observation and the column a variable. The steps taken index is leading in both panels. For an explanation of the relevant variables, please refer to the introduction of Appendix C.

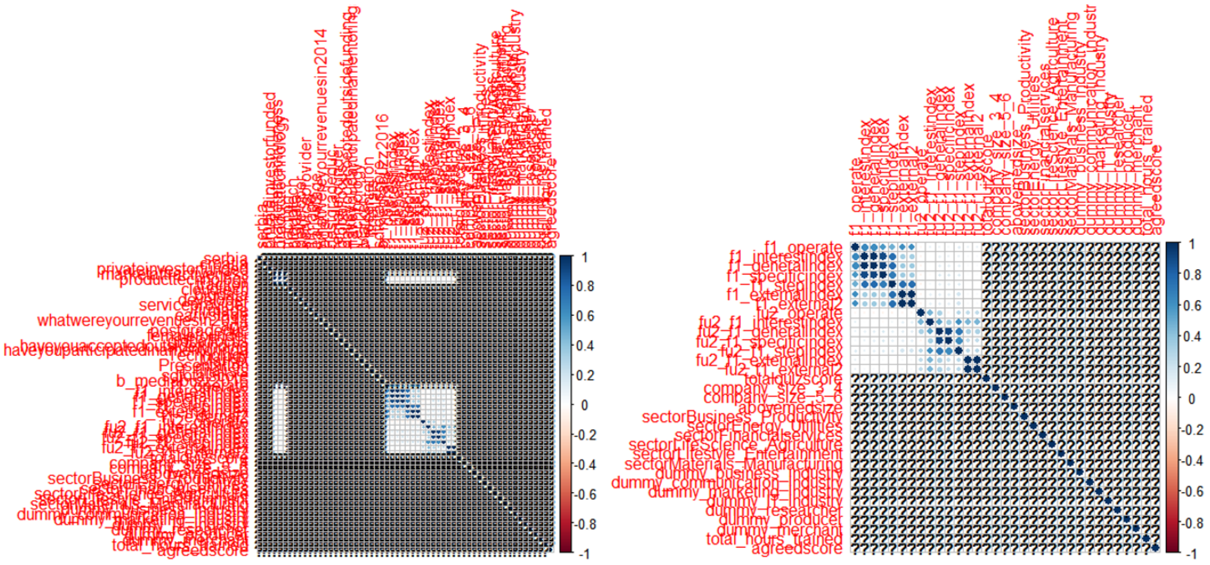

Figure C.3: Missing Value Correlation Diagram

Notes: The left panel shows the correlation between all variables. The right panel zooms in on the part where little correlation is shown. For an explanation of the relevant variables, please refer to the introduction of Appendix C.

C.3 KNN or MICE

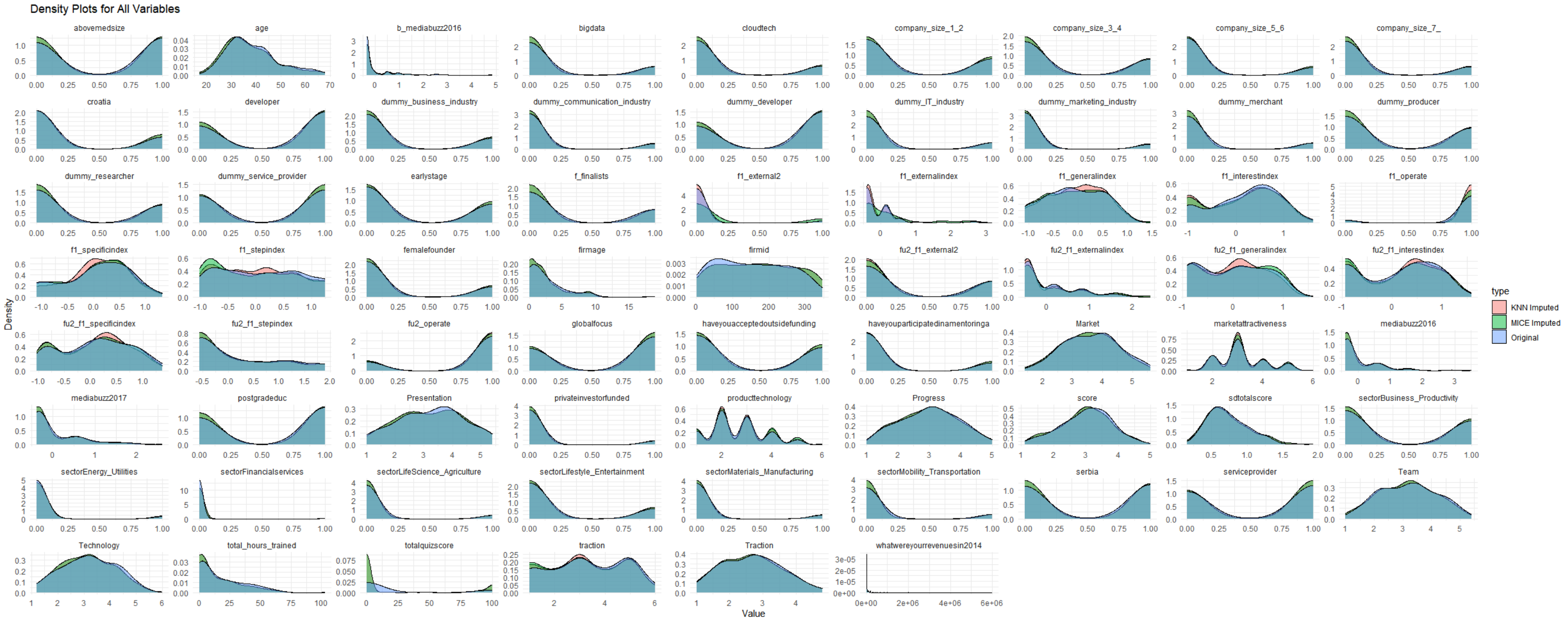

Figure C.4: Density Plots for All Variables with KNN and MICE Imputed Data

Notes: Each panel represents the density plot of a single variable, showing the original data with the kNN and MICE imputed data. For an explanation of the relevant variables, please refer to the introduction of Appendix C.

D Additional Results

In this section of the appendix, an additional table is presented to accompany Figure 4.1 and ??.

Table D.1: GATES of 20% Most and Least Affected Groups

		K-Means			Neural Network	
Impact on Investment Readiness Score	20% Most (G_5)	20% Least (G_1)	Difference	20% Most (G_5)	20% Least (G_1)	Difference
Estimate	0.026	-0.221	0.364	0.116	-0.113	0.025
90% Confidence interval	(-2.087, 2.050)	(-2.427, 1.713)	(-2.346, 2.988)	(-2.469, 2.359)	(-2.921, 2.748)	(-2.985, 3.191)
P-value	0.972	0.796	0.736	0.935	0.924	0.988
		Lasso			Neural Network	
Impact on Making a Deal with an Investor	20% Most (G_5)	20% Least (G_1)	Difference	20% Most (G_5)	20% Least (G_1)	Difference
Estimate	0.007	-0.014	-0.157	-0.028	0.090	-0.127
90% Confidence interval	(-0.956, 0.953)	(-1.280, 1.539)	(-1.828, 1.478)	(-1.225, 1.121)	(-1.272, 1.699)	(-1.871, 1.490)
P-value	0.979	0.962	0.885	0.941	0.890	0.876

Notes: Medians over 100 splits. The Investment Readiness Score is the score given by judges in the semi-finals. The making a deal with an investor is based on the second survey, after two years of the program.

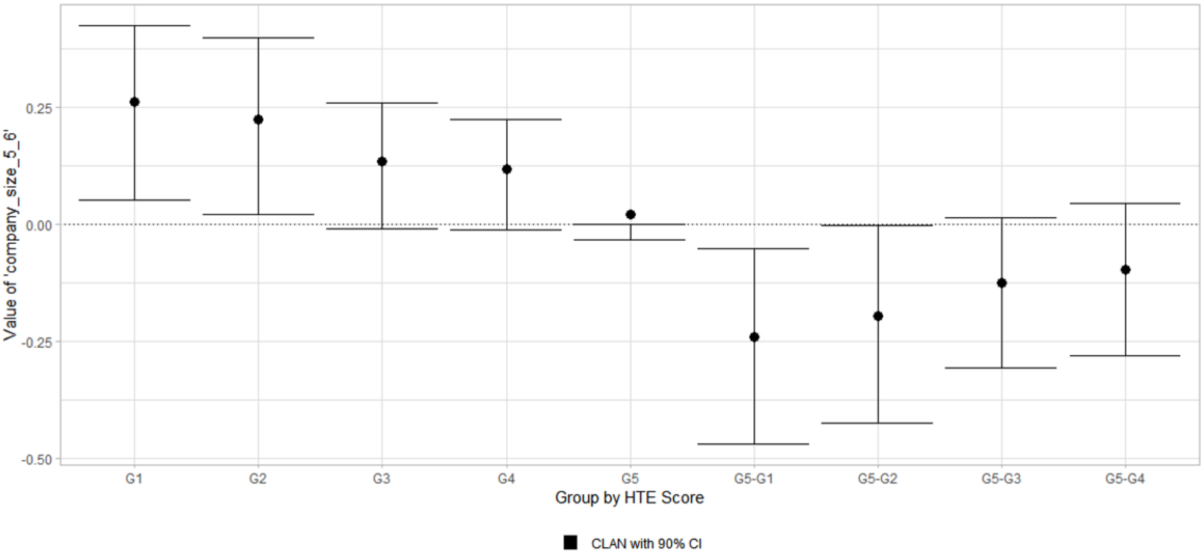

Figure D.1: CLAN on IRS for Company Size: 5-6

Notes: The figure shows the CLAN estimates on IRS per group for the binary variable *company size: 5-6*. The groups $G_1, \ldots, G_5$ each consist of 20% of the data, where G_1 and G_5 are the least and most affected groups, respectively. Furthermore, the variables that follow denote the difference between the most affected group and the other groups. It is notable that for G_5 the estimate lies outside the 90% confidence interval.

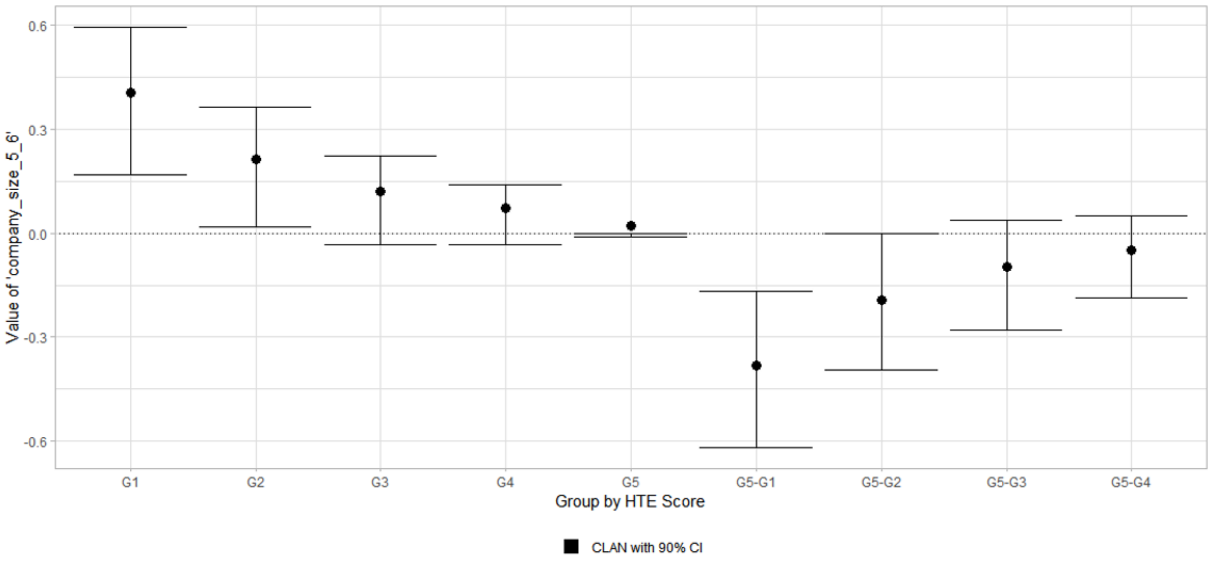

Figure D.2: CLAN on MDI for Company Size: 5-6

Notes: The figure shows the CLAN estimates on MDI per group for the binary variable *company size: 5-6*. The groups $G_1, \ldots, G_5$ each consist of 20% of the data, where G_1 and G_5 are the least and most affected groups, respectively. Furthermore, the variables that follow denote the difference between the most affected group and the other groups. It is notable that for G_5 the estimate lies outside the 90% confidence interval.

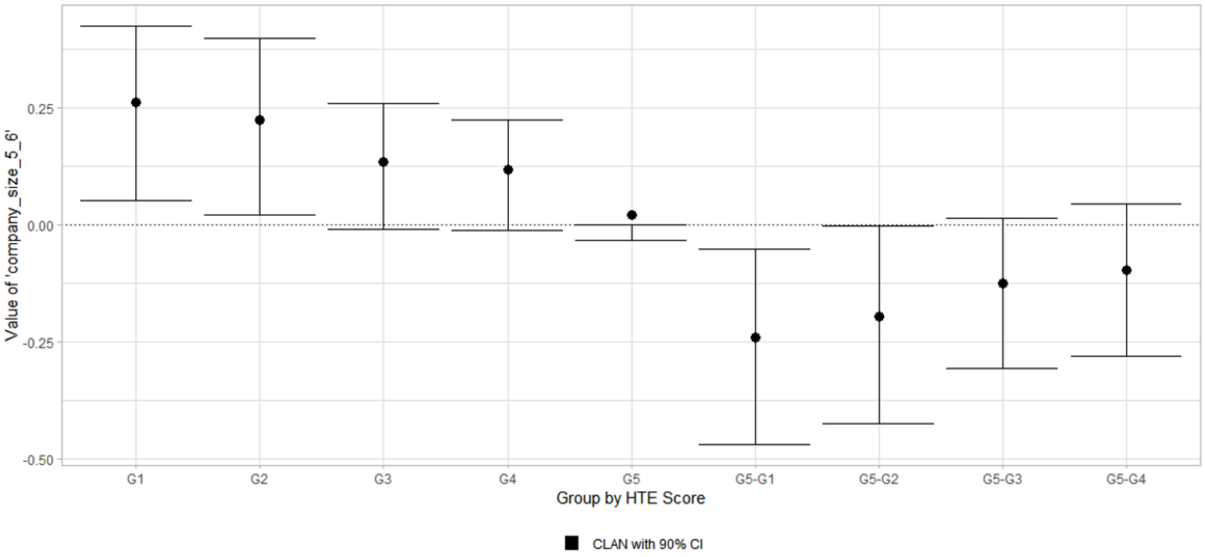

Figure D.3: CLAN on MDI for Industry is Communication & Collaboration

Notes: The figure shows the CLAN estimates on MDI per group for the binary variable *industry is communication & collaboration*. The groups $G_1, \ldots, G_5$ each consist of 20% of the data, where G_1 and G_5 are the least and most affected groups, respectively. Furthermore, the variables that follow denote the difference between the most affected group and the other groups. It is notable that for G_5 the estimate lies outside the 90% confidence interval.

E Replication Microfinancing in Morocco

This section in the appendix contains the replicated tables from the microfinancing result section in Chernozhukov et al. (2017).

Table E.1: Comparison of ML Methods: Microfinance Availability

	Elastic Net	Boosting	Nnet	Random Forest
Amount of Loans				
Best BLP (Λ)	2735486	1998536	2076124	2671565
Best GATES ($\bar{\Lambda}$)	833	409	359	1313
Output				
Best BLP	135341938	85799194	86808131	110644996
Best GATES ($\bar{\Lambda}$)	7920	3727	4518	4853
Profit				
Best BLP	32446221	16941744	16352450	31536586
Best GATES ($\bar{\Lambda}$)	4113	1922	1571	3583
Consumption				
Best BLP	38879	27129	30855	33504
Best GATES ($\bar{\Lambda}$)	91	81	100	96

Notes: Medians over 100 splits.

Table E.2: BLP of Microfinance Availability

	ATE	HET	ATE	HET
Amount of Loans	1089	0.212	1095	0.412
	(528.9,1663)	(0.008,0.451)	(541.6,1666)	(0.001,0.827)
	[0.000]	[0.085]	[0.000]	[0.099]
Output	5365	0.228	4716	0.179
	(539.7,10410)	(0.062,0.391)	(-203.9,9713)	(-0.144,0.451)
	[0.060]	[0.012]	[0.120]	[0.530]
Profit	1372	0.212	1225	0.229
	(-1325,4141)	(0.054,0.396)	(-1442,4069)	(-0.010,0.466)
	[0.633]	[0.019]	[0.701]	[0.119]
Consumption	-57.90	0.130	-67.73	0.161
	(-163.4,47.96)	(-0.084,0.376)	(-176.6,41.89)	(-0.233,0.547)
	[0.547]	[0.383]	[0.491]	[0.823]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

Table E.3: GATES of 20% Most and Least Affected Groups

	Most Affected	Least Affected	Difference	Most Affected	Least Affected	Difference
Amount of Loans	2838	-238.3	3029.0	2860	-205.5	2956.0
	(1441,4297)	(-1797,1249)	(809.8,5176)	(1173,4516)	(-1824,1430)	(504.4,5340)
	[0.000]	[1.000]	[0.014]	[0.002]	[1.000]	[0.042]
Output	20954	-1977.0	22820	20277	784.9	20284
	(6473,36986)	(-12143,7200)	(5356,41033)	(4404,36638)	(-11323,13012)	(-572.5,41291)
	[0.008]	[1.000]	[0.021]	[0.023]	[1.000]	[0.113]
Profit	10539	-924.20	10962	10373	-832.60	11253
	(2357,19130)	(-7134,5299)	(72.22,21929)	(1875,19258)	(-7650,5597)	(-406.1,22815)
	[0.025]	[1.000]	[0.097]	[0.037]	[1.000]	[0.117]
Consumption	34.66	-321.8	341.90	41.17	-274.7	313.50
	(-192.8,265.4)	(-699.8,27.95)	(-78.03,818.5)	(-258.9,334.9)	(-667.3,104.2)	(-224.2,847.9)
	[1.000]	[0.139]	[0.225]	[1.000]	[0.303]	[0.493]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

Table E.4: CLAN of Microfinance Availability

	Most Affected	Least Affected	Difference	Most Affected	Least Affected	Difference
Amount of Loans						
Head Age	31.20	39.30	-8.272	24.85	39.03	-14.280
	(29.12,33.35)	(37.22,41.43)	(-11.26,-5.207)	(22.76,26.94)	(36.92,41.21)	(-17.28,-11.27)
	-	-	[0.000]	-	-	[0.000]
Number of Household Members	3.361	4.569	-1.244	2.608	4.541	-1.903
	(3.083,3.644)	(4.291,4.847)	(-1.639,-0.849)	(2.326,2.885)	(4.261,4.822)	(-2.300,-1.516)
	-	-	[0.000]	-	-	[0.000]
Total Borrowed	0.167	0.221	-0.059	0.131	0.279	-0.148
	(0.133,0.200)	(0.188,0.253)	(-0.106,-0.012)	(0.097,0.164)	(0.245,0.313)	(-0.196,-0.099)
	-	-	[0.026]	-	-	[0.000]
Output						
Head Age	36.57	35.75	0.608	35.47	31.20	3.494
	(34.45,38.67)	(33.66,37.88)	(-2.400,3.633)	(33.25,37.69)	(28.97,33.43)	(0.345,6.644)
	-	-	[1.000]	-	-	[0.060]
Number of Household Members	4.070	3.767	0.299	3.860	3.481	0.462
	(3.791,4.348)	(3.491,4.045)	(-0.093,0.698)	(3.565,4.155)	(3.184,3.780)	(0.045,0.878)
	-	-	[0.268]	-	-	[0.060]
Total Borrowed	0.186	0.237	-0.050	0.200	0.193	0.010
	(0.151,0.221)	(0.203,0.273)	(-0.098,-0.002)	(0.165,0.233)	(0.160,0.225)	(-0.038,0.056)
	-	-	[0.086]	-	-	[1.000]
Profit						
Head Age	34.80	38.47	-3.966	32.58	34.27	-1.945
	(32.71,36.89)	(36.42,40.53)	(-6.859,-0.931)	(30.34,34.87)	(31.98,36.56)	(-5.198,1.256)
	-	-	[0.021]	-	-	[0.468]
Number of Household Members	3.942	3.893	0.040	3.479	3.744	-0.309
	(3.661,4.227)	(3.619,4.168)	(-0.346,0.415)	(3.187,3.769)	(3.449,4.035)	(-0.712,0.095)
	-	-	[1.000]	-	-	[0.276]
Total Borrowed	0.180	0.253	-0.080	0.158	0.187	-0.029
	(0.145,0.214)	(0.219,0.287)	(-0.129,-0.033)	(0.127,0.189)	(0.156,0.218)	(-0.074,0.016)
	-	-	[0.002]	-	-	[0.392]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
P-values for the hypothesis that the parameter is equal to zero in brackets.

F Replication Western Balkans

This section in the appendix contains the replicated figures and tables presented in Cusolito et al. (2021).

The variables in the figures and tables below are defined as follows. **Media Buzz** is a standardized index of whether the firm is mentioned in the media, the number of media mentions, number of Facebook likes and number of Twitter followers. **Firm survival** is a binary variable that takes value one if the firm is operating, and zero otherwise. **Interested in equity** is a standardized index of whether the firm is interested in equity financing, the maximum equity share they are willing to have owned by outside investors, whether they have specific deal terms for investors, and whether they would consider a royalty-based investment. **General investability** is a standardized index of number of employees, whether the founders work full-time in the business, whether the firm had positive sales in the first quarter of the year, whether total sales exceed 10,000 euros in that quarter, whether the firm made a positive profit in the past year, and whether the firm made sales to Western Europe or the United States. **Specific needs of investors** is a standardized index of whether business and personal accounts are separated, whether the firm has made a revenue projection for the next year, whether it knows customer acquisition costs, the number of key metrics tracked, whether it has found out if the product or service can be covered by intellectual property protection, and whether it has at least one form of intellectual property protection received or pending. **Investment steps** is a standardized index of having contacted an outside investor, made a pitch to an outside investor, have a mentor or external expert supporting them to obtain financing, and entered into negotiations with an outside investor. **External investment** is a standardized index of having taken on new debt, having made a deal with an outside investor, have received at least 25,000 euros in outside financing, and have received an incubator or accelerator grant (all since August 2015). **Made a deal with an investor** indicates having made a deal with an outside investor since August 2015 (program start).

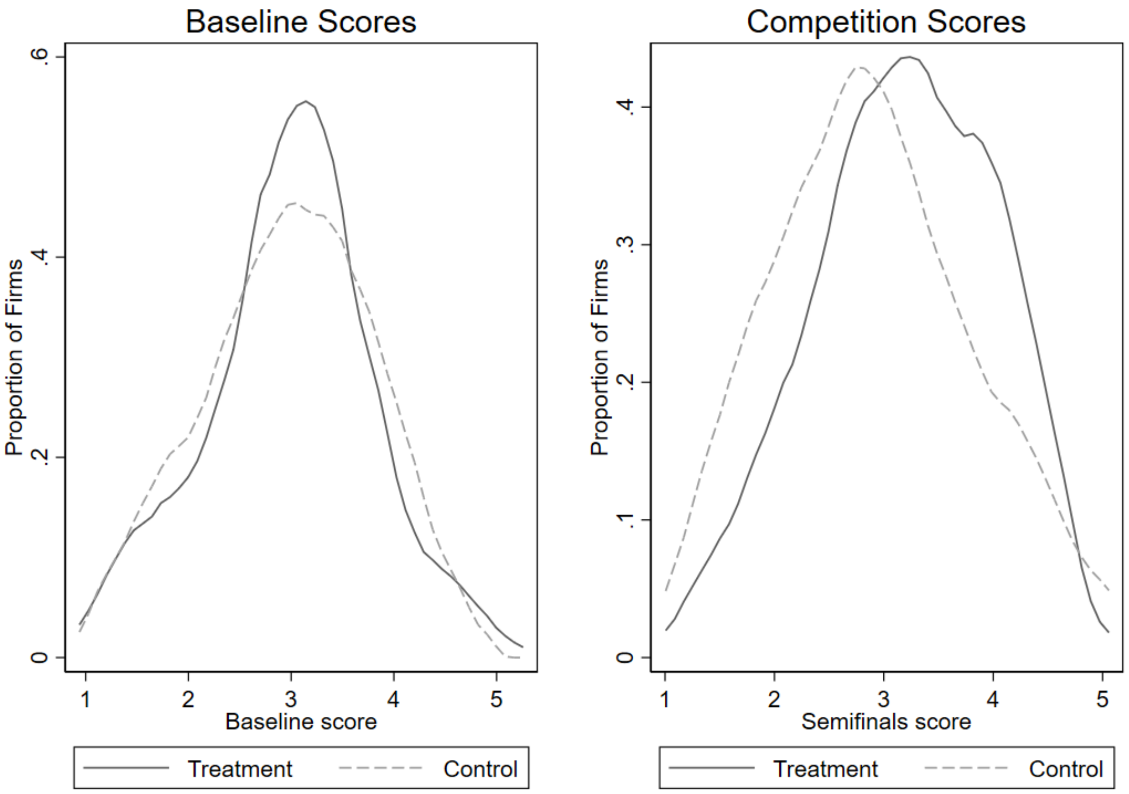

Figure F.1: Distributions of Baseline and Post-Intervention Competition Investment Readiness Scores by Treatment Status

Notes: Baseline scores are for the subset of firms that attended the semi-finals. Competition scores are post-treatment. Kolmogorov-Smirnov test of equality of distributions has p-value of 0.959 at baseline and 0.017 post-intervention.

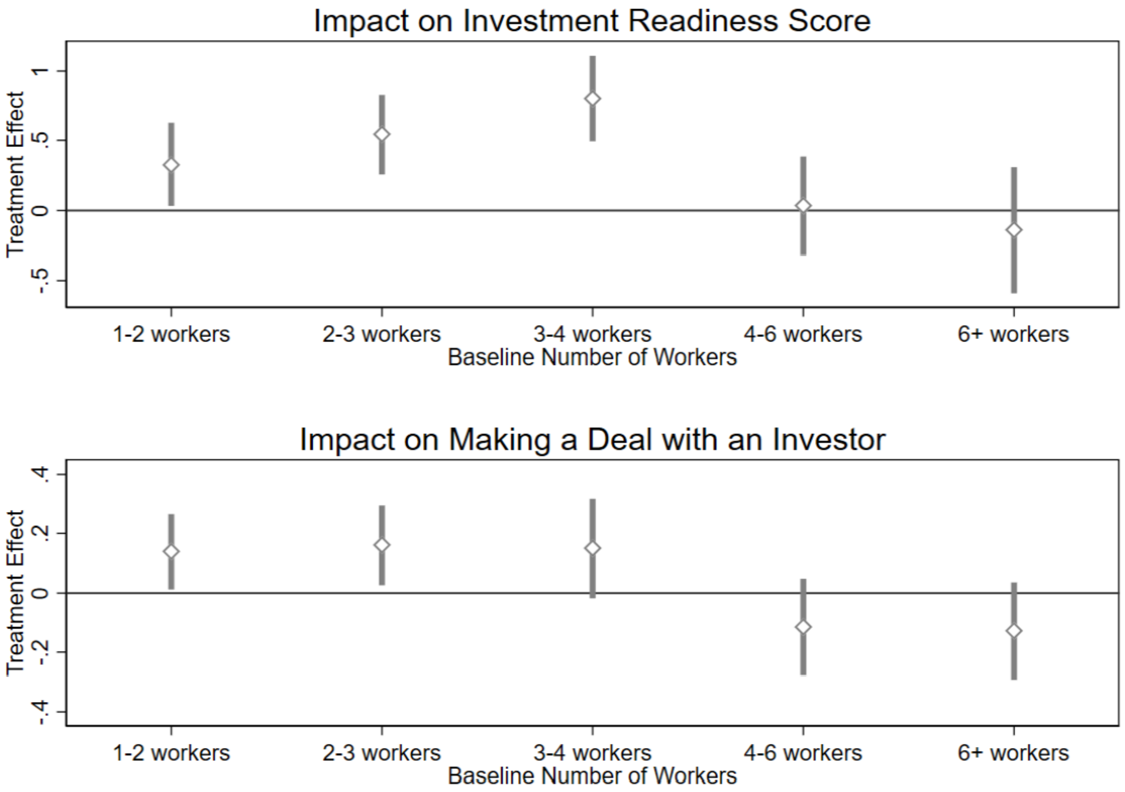

Figure F.2: The Impact of the Program is Higher for Firms that Were Small at Baseline

Notes: Graphs show point estimates from rolling regressions which estimate the impact of being assigned to treatment for rolling samples of approximately 30 percent of the sample at the time, conditioning on the stratifying variables of initial investment readiness score, country, and whether or not the firm had a private investor to begin with. 90 percent confidence intervals shown around point estimates.

Table F.1: Impact on Program on Investment Readiness as Scored by Judges

	Pitch score: Overall score	Pitch score: Team	Pitch score: Technology	Pitch score: Traction	Pitch score: Market Attractiveness	Pitch score: Progress	Pitch score: Presentation	Selected to go to Finals	Standard deviation of Score amongst Judges
Base Specification									
Assigned to Treatment	0.284**	0.167	0.372**	0.206	0.268*	0.373***	0.372**	0.115*	0.006
	(0.126)	(0.150)	(0.152)	(0.130)	(0.137)	(0.137)	(0.164)	(0.068)	(0.049)
Including Judge Fixed Effects									
Assigned to Treatment	0.409***	0.369**	0.476***	0.295**	0.463***	0.440***	0.514***	0.090	-0.017
	(0.135)	(0.158)	(0.174)	(0.142)	(0.139)	(0.143)	(0.191)	(0.076)	(0.051)
Sample Size	211	211	211	211	211	211	211	211	211
Control Mean	2.908	3.042	2.970	2.541	3.406	2.794	3.042	0.122	0.723
Control Std. Dev	0.903	1.068	1.031	0.947	0.940	0.937	1.145	0.328	0.317

Notes: Robust standard errors in parentheses. Regressions control for randomization strata. *, **, *** indicate significance at the 10, 5, and 1 percent levels respectively. The judge fixed effects controls for which five of the sixty-five judges judged a particular firm.

Table F.2: Impacts on Firm Outcomes 6 Months and 2 Years after Program

	Media Buzz	Firm Survival	Interest in Equity	General Investability	Specific Needs of Investors	Investment Steps Investment Steps	External Investment	Made Deal with Investor
Panel A: Impact at Six Months								
Assigned to treatment	0.085	0.049	0.051	0.026	0.082	-0.017	-0.152*	-0.024
	(0.053)	(0.030)	(0.094)	(0.085)	(0.080)	(0.098)	(0.087)	(0.033)
Sample Size	346	319	278	277	269	240	279	279
Control Mean	-0.060	0.898	-0.015	-0.039	-0.059	0.008	0.084	0.083
Control S.D.	0.546	0.303	0.764	0.634	0.682	0.720	0.741	0.276
Panel B: Impact at Two Years								
Assigned to treatment	0.112**	0.072	0.032	0.089	0.084	0.044	0.003	0.050
	(0.047)	(0.045)	(0.084)	(0.082)	(0.079)	(0.092)	(0.080)	(0.049)
Sample Size	346	340	309	291	298	282	330	330
Control Mean	-0.073	0.753	-0.003	-0.057	-0.058	-0.033	0.021	0.244
Control S.D.	0.528	0.433	0.782	0.649	0.692	0.760	0.700	0.431

Notes: robust standard errors in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent levels respectively. All regressions control for randomization strata fixed effects.

Table F.3: Judges Scores Predict Firm Outcomes 6 Months and 2 Years after Program

	Media Buzz	Firm Survival	Interested in Equity	General Investability	Specific Needs of Investors	Investment Steps	External Investment	Made a Deal with investor
Panel A: Association at Six Months								
without controls								
Score assessed by Judges	0.261***	0.024	0.201**	0.076	0.336***	0.222***	0.213**	0.093**
	(0.057)	(0.037)	(0.076)	(0.072)	(0.065)	(0.082)	(0.098)	(0.038)
with controls for country, prior funding, sector, firm age and stage, founder gender and education, baseline employment								
Score assessed by Judges	0.220***	0.017	0.209**	0.080	0.296***	0.155	0.190*	0.080**
	(0.052)	(0.042)	(0.099)	(0.080)	(0.078)	(0.103)	(0.113)	(0.037)
Sample Size	101	92	83	83	81	73	82	82
Control Mean	-0.060	0.898	-0.015	-0.039	-0.059	0.008	0.084	0.083
Control S.D.	0.546	0.303	0.764	0.634	0.682	0.720	0.741	0.276
Panel B: Association at Two Years								
without controls								
Score assessed by Judges	0.271***	0.061	0.153*	0.040	0.136*	0.322***	0.324***	0.166***
	(0.059)	(0.041)	(0.088)	(0.073)	(0.082)	(0.100)	(0.072)	(0.048)
with controls for country, prior funding, sector, firm age and stage, founder gender and education, baseline employment								
Score assessed by Judges	0.232***	0.079	0.153	0.061	0.128	0.331***	0.334***	0.163***
	(0.053)	(0.049)	(0.097)	(0.082)	(0.086)	(0.103)	(0.082)	(0.051)
Sample Size	101	100	92	86	88	80	99	99
Control Mean	-0.073	0.753	-0.003	-0.057	-0.058	-0.033	0.021	0.244
Control S.D.	0.528	0.433	0.782	0.649	0.692	0.760	0.700	0.431

Notes: Robust standard errors in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent levels respectively.

Predicted Treatment effect is the treatment effect predicted from the association in the control group between the judges' score and the outcome, multiplied by the treatment effect of the program on the judges' score. Outcomes are as defined in Table F.2.

Table F.4: Heterogeneity in Impacts by Initial Firm Size

	Investment Readiness Score	Media Buzz	Firm Survival	Interested in Equity	General Investability	Specific needs of Investors	Investment Steps	External Investment	Made a Deal with Investor
Assigned to Treatment	0.474***	0.157**	0.112	0.071	0.034	0.081	0.217	0.194*	0.156**
	(0.172)	(0.080)	(0.071)	(0.127)	(0.112)	(0.122)	(0.142)	(0.108)	(0.071)
Treatment * Company Size: 4+	-0.450	-0.056	-0.087	-0.073	0.110	0.005	-0.338*	-0.390**	-0.212*
	(0.291)	(0.133)	(0.108)	(0.193)	(0.188)	(0.187)	(0.203)	(0.193)	(0.113)
Company Size: 4+	0.422*	0.184*	0.145*	-0.001	0.250*	0.159	0.135	0.373***	0.165**
	(0.216)	(0.098)	(0.081)	(0.153)	(0.136)	(0.144)	(0.163)	(0.126)	(0.081)
Sample Size	211	346	340	309	291	298	282	330	330
Control Mean Small Firms	2.683	-0.207	0.701	-0.051	-0.216	-0.139	-0.110	-0.171	0.143
Control Mean Larger Firms	3.198	0.065	0.807	0.046	0.106	0.023	0.044	0.223	0.350

Notes: Median Size or higher is a dummy variable taking value one if the firm has at least the median number of baseline workers, and zero otherwise. Investment Readiness Score is Score as assessed by Judges. Robust Standard Errors in parentheses, *, **, and *** denote significance at the 10, 5, and 1 percent levels respectively. Outcomes are taken from two year follow-up survey, and are two years post-intervention.

Table F.5: Heterogeneity in Treatment by Predicted Likelihood of Making a Deal with an Investor

	Impact on Making a Deal within 2 Years	
	Leave-one-out Estimator	Repeated Split-Sample Estimator
Low Predicted Likelihood of Funding	0.143**	0.124**
	(0.066)	(0.055)
High Predicted Likelihood of Funding	-0.081	0.049
	(0.079)	(0.074)

Notes: Bootstrap standard errors, based on 500 bootstrap replications, are reported in parentheses.

*, **, *** denote significance at the 10, 5, and 1 percent levels respectively.

The repeated split-sample estimator uses 200 splits of the data.

Predicted likelihood of funding based on the following baseline characteristics:

Employment above the median, initial investment readiness score, country, whether the firm has had a private investor, whether it classifies itself as early stage, sector, firm age, whether the main founder has post-graduate education, whether at least one founder is female, and whether the firm has previously received mentoring. Abadie et al. (2018) endogenous stratification approach used.

G Programming

For the replication of the results of micro-financing in Morocco, please run the *Thesis Replication Code* → *Morocco* → *Micro in Morocco* file in R. Please adjust the path of the source and the *data12345* variable accordingly.

For the replication of the results of the Western Balkans, please run the *Thesis Replication Code* → *Western Balkans* → *ReplicateTablesandFigures* file in Stata. Please adjust the path of the directory accordingly.

For the replication of the application of the generic ML approach and all remaining results in this paper, please run the *Thesis Replication Code* → *GenericML Application Western Balkans* → *Extension_Western_Balkans* file in R.