CERGE - EI

Centre for Economic Research and Graduate Education – Economics Institute

Charles University

Revisiting Treatment Effects with Causal Forests

Aslan Bakirov

Master Thesis

Prague, July 2023

I declare that I carried out this master's thesis independently, and only with the cited sources, literature and other professional sources. I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In ..... date ..... signature of the author

Acknowledgement

I would like to express my gratitude to my supervisor, Paolo Zacchia, for his constant encouragement and honest feedback. I also extend my thankfulness to our faculty members with whom I had discussed parts of this work, and the ASC for motivating me to be a better writer. Last but not least, I thank my friends and colleagues for their camaraderie and support.

Abstract

This thesis focuses on the application of Causal Forests, a prominent causal machine learning algorithm, to estimate heterogeneous treatment effects in complex socio-economic phenomenon. Causal Forests leverage the capabilities of random forests to partition the high-dimensional covariate space and identify subgroups where the effect of an intervention remains constant. This approach is particularly valuable when dealing with heterogeneous causal effects, where a uniform measure of gains for all is an unrealistic assumption. Unlike traditional manual methods that are susceptible to p-hacking, the algorithm objectively uncovers nuanced treatment effect variations through data-driven analysis. The thesis demonstrates the algorithm's potential in exploring causal effects and providing valuable policy insights. An empirical illustration showcases the modeling of a complex socio-economic phenomenon, such as the gender wage gap, and leverages Causal Forests to extract policy learning from the identified heterogeneity. The study highlights the algorithm's contribution to credible and robust causal inference, bridging the gap between traditional decomposition methods and data-informed heterogeneity analysis.

Keywords: Causal machine learning, heterogeneity, policy learning, policy targeting, gender gap

Contents

1 Introduction

Since the 1980s, applied economic research has been increasingly focused on achieving clearer causal understanding by emphasizing credible identification of parameters. Micro-level questions gained prominence as traditional aggregate models struggled to establish definitive causal links, especially concerning inflation and unemployment. The development of human capital theory, rebirth of institutionalism, efforts to address poverty and unemployment, among others, further contributed to the demand for research that could offer more credible inferences (Angrist and Pischke 2010).

The evolution of the quality empirical research in applied work covered both model-based and design-based estimation. The availability of data played a significant role too, since earlier on the aggregate data provided less support for micro-level variation in the models, thus resulting in unclear inference. As more granular data became available, the methods based on that also added in credibility.

The demand was driven in part by the policy evaluation framework too. After the stagflation of 1970s, the US government, among others, launched several job training programs aimed at reducing the unemployment. At the time it was known that the ideal way to approach the questions of cause was with randomization, but the widespread adoption of this was lagging (Angrist and Pischke 2010; Lechner 2023). Although the idea of randomization-based inference existed since at least the 1920s, it was formalized in the **potential outcomes framework** by Rubin (1974), leading to increased adoption of credible policy evaluation practices.

Randomized Controlled Trials (RCTs) were (are) a powerful tool, but expensive and sometimes impractical in economics. In his critique, Leamer (1983) not only pinpointed the poor quality of data analysis, but also suggested that one should report a series of stress tests to establish robustness of the parameters when the ideal RCT is unavailable. However, methods that exploited quasi-experimental design like Instrumental variables (IV), differences-in-differences (DiD) and regression discontinuity designs (RDD) added in volume. Alongside the change in the understanding of basic regression analysis and matching, the design-based methodology proved to be more useful, more beloved, and more presentable to the general public (Angrist and Pischke 2010; Lechner 2023). The potential outcomes framework emerged as a pivotal tool for formulating causal links,

independent of model specifications. In this thesis, I mainly follow potential-outcomes notations, though it is not the only framework to explore causality. More recently Judea Pearl, a renowned computer scientist developed his own framework with Directed Acyclic Graphs (DAGs). DAGs are based on the graph theory and can be more visually appealing than potential outcomes, yet the two are very similar in the mechanics (Pearl et al. 2000).

The improvement in the quality of applied econometric work was accompanied by advancements in computational resources. With increased computing power, firms in private sector with substantial capital could collect and store larger datasets. Recognizing the importance of random experimentation, some of these firms conducted A/B tests on a larger scale and with higher frequency. For instance, Microsoft conducted thousands of A/B tests annually (Cunningham 2022b). Meanwhile, because of the exceptional accuracy in prediction tasks, machine learning methods have been a preferred choice for decision-making in most cases. This is not unexpected, since machine learning can be seen as a more practice-oriented integration of statistics and computer science.

Yet the last decade or so saw visible changes in interest for causal identification in the private sector too, particularly in the big corporations. While some questions required structural modeling from microeconomic perspective (auction design, for instance), even more focused on evaluation of causal effects of interventions and targeting (Athey 2017). Combined with the data-rich setting, this trend has given rise to the fusion of causal inference and machine learning. Economists working closely with industry have attempted to adopt the machine learning methods to potential outcomes, or prediction for causal inference purposes, and as an important step derived important results to establish trust in those "black-box" algorithms. Causal Forest is a prominent example of causal machine learning algorithms, initially proposed by Wager and Athey (2018). It capitalizes on the capabilities of random forests (Breiman 2001) to partition the covariate space, identifying subgroups where the intervention's effect remains constant. This proves particularly valuable when examining causal effects with heterogeneity, where a uniform measure of benefit cannot be prescribed to the entire sample. Estimating such heterogeneous treatment effects demands meticulous development of covariates and thorough testing of all interactions, which can be a tedious process.

In practical research, it is essential for researchers to develop a pre-analysis plan,

specifying hypothesized subgroups where the treatment effect may be stronger or weaker. By doing so, researchers proactively outline their analysis approach and hypotheses before conducting the actual exploration. This pre-analysis plan serves as a protective measure against the risk of "p-hacking" or "data-mining", where researchers might manipulatively test potential subgroups until they find significant variations in treatment effects. Such practices compromise the integrity of academic research, making it vulnerable to manipulation and false discoveries (Athey and Imbens 2019). Causal forests and similar meta-algorithms can offer an objective approach to identify the subgroups with the strongest or weakest treatment effects, guarding against manipulation and promoting more robust research.

In this thesis, I intend to contribute to the growing strands of literature on causal machine learning, and gender wage gap decomposition. I begin by reviewing the fundamental concepts of causal inference, focusing on the potential outcomes framework with observational data. I briefly summarize the assumptions required to recover causal effects in constant effects scenario, and then under heterogeneity, aiming to understand how machine learning models can be adapted for causal inference.

Next, I provide a brief overview of machine learning methods and relevant algorithms, highlighting their differences and similarities compared to traditional econometric modeling. I then delve into the recent emergence of causal machine learning, discussing main approaches and results, along with illustrative examples such as prediction policy problems, post-selection inference, and doubly-robust estimation.

As an empirical illustration, I revisit the analysis of Huber and Solovyeva (2020) of the gender wage gap in the U.S. around the year 2000. They use the DAG framework to formulate the estimation of the gap, and its further decomposition into direct (discrimination) and indirect (mediated) effect (Huber and Solovyeva 2020). While their analysis provides valuable insights into the causal effects of gender perception on wages, it also highlights the sensitivity of the estimates due to the modeling choices (Huber and Solovyeva 2020). As a natural next step, a more policy-relevant approach involves identifying subgroups of women facing lower pay compared to their male counterparts. Modeling of a social phenomenon as this with full consideration of the possible covariate sets requires clear emphasis on the multiple hypotheses testing in high-dimensional space. Using Causal Forests, I explore heterogeneity across various covariates to uncover

these subgroups and showcase the algorithm's power in capturing nuanced differences within the gender wage gap.

Finally, I conclude my thesis, summarizing the findings and contributions to the field of causal inference and gender wage gap decomposition, as well as applications for causal machine learning.

It is important to note however that I aim not to add innovation to the methodology of Causal Forest, but to utilize it. Nor do I focus on other methods applicable in this setup. The setup I follow only includes a treatment variable which is binary, and my identification strategy is identical to that of Huber and Solovyeva (2020). Using their data and identification, I intend to show how one can apply Causal Forests to explore heterogeneity in the gender wage gap.

2 Literature review

2.1 Hindrances to Exploring Causal Effects

Papers focusing on deriving causal inference nearly always rely on the potential outcomes framework of Neyman, Rosenbaum, Rubin and others. As per the seminal formulation put forth by Neyman (1923), the causal effect of an intervention is conventionally viewed as the difference between the observed outcome and the hypothetical outcome that would have transpired in the absence of the said intervention. This delineation effectively lays the groundwork for the concept of counterfactuals, which has been subsequently formalized into the comprehensive potential outcomes framework by Rubin (1974) and has since become integral to causal inference research. When conducting an experiment with an intervention W, we can have the outcomes measured for control and treatment groups as Y^0 *and* Y^1, respectively. However, to draw a conclusion that the intervention caused the difference in outcomes, we need more than just these measurements. One needs a well-designed study, appropriate statistical methods, control of other factors that could influence the results, and ideally, random assignment of participants to the treatment and control groups. These elements are essential for making valid causal inferences and establishing a strong case for the intervention's actual impact on the

outcomes. The effect of the intervention for an individual i, *ceteris paribus*, is then

$$TE = Y_i^1 - Y_i^0$$

or, as one is mostly interested in average of that for the whole sample:

$$ATE = E[Y_i^1 - Y_i^0]$$

This expression demands that we observe both states of an individual, treated and not, and then measure the difference for the person. This is an issue that we would not be able to address even if we had access to an infinite amount of data, making it the fundamental problem of causal inference (Holland 1986). Specifically, we face the obstacle of not being able to observe the "counterfactual" state of an individual under a different treatment rule. In other words, we cannot simultaneously see what would have happened if the unit received the treatment and what would have occurred if they did not. To emphasize, we cannot observe the **counterfactual** of an individual under different treatment rule. Thus we do not have access to the **"ground truth"** to evaluate our attempts estimating the average treatment effect (ATE) either.

2.1.1 Selection Bias

This is important since one might be tempted to estimate the true ATE as a simple difference in means of the two group outcomes $\hat{ATE}$, and that arises the question of selection bias:

$$\begin{aligned} \hat{ATE} = E[Y_i^1 - Y_i^0] = E[Y_i^1|W = 1] - E[Y_i^0|W = 1] + \\ + E[Y_i^1|W = 0] - E[Y_i^0|W = 0] = \\ = E[Y_i^1|W = 1] - E[Y_i^0|W = 0] + bias \end{aligned}$$

However, as mentioned in the previous section, Neyman and Fischer in 1920s had already shown that the physical randomization of the treatment assignment helps identify the ATE by alleviating the selection bias. As Rubin integrated this idea into the potential outcomes framework, we can formulate the need for randomization further:

$$Y_i^1, Y_i^0 \perp W_i$$

In other words, the potential outcomes in two states should be independent of what an individual's treatment status is. This assumption allows the researcher to make sure there is no self-selection induced by the individuals' expectation of their gain/ loss from the intervention. Although, random assignment mechanism works wonders, it is not always cost-efficient, and sometimes downright impossible to randomize anything. A natural next step is to try and replicate the results of the randomization with observational data. However, when working with observational data, it is critical to ensure that there is no self-selection into or out of treatment.

For addressing this issue, critical assumptions have to be made. Particularly,

- **the assumption of unconfoundedness + ignorability**;

  Ignorability ensures that selection into treatment is independent of the potential outcomes, while unconfoundedness requires that no factor (observed or unobserved) interferes with the analysis. In a randomized controlled trial both follow from design. Although mathematically identical, the terms ignorability, unconfoundedness, and even conditional independence (CIA) can be used interchangeably, but one can see the complementarity between them. One cannot obtain unconfoundedness without ignorability, and ignorability without unconfoundedness. To formulate it in the potential outcomes framework, let us compose a set of characteristics X_i that is sufficiently rich to account for any kind of personal motivation to self-select into the treatment. Then,

$$Y_i^1, Y_i^0 \perp W_i | X_i$$

That is, we need to control for the X_i in our outcome equation, as in the following:

$$\begin{gathered} E[Y_i|W_i, X_i] = E[Y_i^0|X_i, W = 0] + \\ + E[\ (E[Y_i^1|X_i, W = 1] - E[Y_i^0|X_i, W = 0])|\ X_i]\ W_i + \epsilon_i = \\ \mu(0, X_i) + E[\mu(1, X_i) - \mu(0, X_i)]\ W_i + \epsilon_i \end{gathered}$$

where $\mu(W_i, X_i) = E[Y|X_i, W_i = w]$ is a flexible control function estimator for Conditional Expectation Functions (CEFs). Partialling out the X_i allows to level the variation

in the outcome between the control and treatment, shrinking the bias in parameter of interest ATE. LaLonde (1986) tried to see if one could replicate the results of random assignment with the control function approach. He started with the job training programs of 1970s where the participants were randomized into treatment originally, hence he knew what to benchmark against. Keeping the original treatment group, he composed six control samples that were different from the original one. He addressed the selection into treatment using the two-step correction method as in Heckman (1979). The results, apart from wide differences in parameter values, show that control functions may not always account for selection bias, especially if the samples are incomparable in the first place. It is also essential to achieve a good fit with the $\mu(W_i, X_i)$ in the outcome equation specified above (LaLonde 1986).

Considering that control functions do not make the units in both groups more comparable, only normalize the variation in the outcome, we need to make sure there are enough points of comparison for each cluster of X_i values in treatment and control groups. This can be achieved with reweighing, particularly on the **propensity score** (PS). The premise of PS is to compose an index out of those X_is and reweigh the samples to ensure covariate balance in the two groups.

$$e(X_i) = E[W_i|X_i] = Pr(W = 1|X_i)$$

Where the $e(X_i)$ is the propensity score that acts as a proxy for probability to select into treatment. This requires the **unconfoundedness** and

- **the overlap assumption (common support)**;

For every value of that index of individual characteristics, there should be at least one matching person in the counter-group. To measure the uncertainty better, it is desired to have more than one unit that is comparable. That is,

$$0 < P(W_i = 1|X_i) < 1 \ \forall i$$

Reflecting on the job training programs evaluation, Smith and Todd (2005) revisit the results of LaLonde (1986) with matching. Their results look better with the original estimates from randomization. Yet, even after ensuring the common support, some selection bias persists. Generally, as Heckman et al. (1998) show it is extremely hard

to get rid of the selection bias without randomization. The control functions and PS matching when done properly reduce it, yet do not fully eliminate. Methods exist that impose parametric assumptions on errors of a model as in Heckman (1979) or Roy (1951), or are non-parametric for high-dimensional settings such as Arellano and Bonhomme (2017) or Chernozhukov, Fernández-Val, and Luo (2023), but those are not within the scope of this thesis.

Another critical assumption that is usually understated, yet plagues both randomization results and observational studies, is that there are no spillovers. In potential outcomes framework it is known as

- **the stable unit treatment value assumption (SUTVA)**;

requiring that nobody's outcome is affected by allocation status of the others. SUTVA fences against inter-group spillovers of the particular intervention.

With the aforementioned assumptions in mind, one can approach the assessment of the causal effect of an intervention by reformulating it as follows:

$$\tau = E[Y_i^1] - E[Y_i^0] = E[\ W_i\ Y_i^1 - (1 - W_i)Y_i^0\ ]$$

where W_i is the binary treatment status of an individual i. For observational studies, we can invoke the unconfoundedness and overlap assumptions to arrive at the **Inverse Propensity Weighing** (IPW) estimator this can be modified as follows:

$$\tau_{IPW} = E[Y_i^1] - E[Y_i^0] = E[\ W_i\ Y_i^1 - (1 - W_i)Y_i^0\ ] = E\left[\frac{W_i\ Y_i}{e(X_i)} - \frac{(1 - W_i)\ Y_i}{1 - e(X_i)}\right]$$

This notation makes use of the propensity score of an individual, $e_i(X_i)$, meaning how likely is that this particular individual receives intervention, based on the covariate values X_i. Weighing by the inverse of the propensity score allows one to correct for the selection into the treatment, achieving covariate balance between treatment and control groups. This provides a way to make units in the two groups more comparable.

It is important to note that, both control function approach and the IPW are sensitive to the choice of X_i and the functional form. In IPW we heavily rely on the accuracy of our non-parametric estimates $\hat{e}(X_i)$, while with control functions it is both $\hat{\mu}(0, X_i)$ and $\hat{\mu}(1, X_i)$. It is only natural to ponder if the two can be combined. In

fact, as Robins, Rotnitzky, and Zhao (1994) first present it, the combined estimator is extremely flexible, is less prone to inconsistency, and is semiparametrically efficient. The core concept is to make a sincere attempt to provide the most accurate estimate of the CEFs $\mu(W_i, X_i)$, and then to recover the unexplained part with the IPW with even the least accurate $e(X_i)$. We arrive at the new **Augmented IPW** (AIPW) estimator τ_{AIPW}:

$$\tau_{AIPW} = E\left[\frac{W_i\ (Y_i - \mu(1,X))}{e(X_i)} - \frac{(1-W_i)\ (Y_i - \mu(0,X))}{1-e(X_i)}\right]$$
$$-E[\ \mu(1,X) - \mu(0,X)\ ]$$

We can see how the flexible-score logic holds if we start with the good enough PS $e(X_i)$ too, since the leftover error is approximately zero, and hence recycles even the most "garbage" estimates for CEFs $\mu(W_i, X_i)$. This is the the so-called **doubly-robustness** property, and the AIPW is sometimes referred to as "the doubly-robust score". Indeed, this method allows for a greater margin for error, as the estimator is consistent when either one of the nuisance parameters (propensity score $e(X_i)$ or CEF $\mu(W_i, X_i)$) converges (I expand on this in Section 2.3.2). As Chernozhukov et al. (2018) show, the doubly-robust approach reduces bias in estimator greatly even when the CEF is misspecified, which is most likely the case with big-data settings. That is because the nuisance parameters are secondary, and we are not interested in their causal inference, but need them to derive inference for the treatment parameter. Hence we can treat their estimation as a prediction task and employ any "black-box" machine learning method that is suitable, and it will not compromise the credibility of the research question. In fact, quite the opposite, we add credibility by abstaining from parametric specifications. Furthermore, the AIPW has the property of being the optimal one in the class of non-parametric estimators, attaining the bound for semiparametric efficiency (Hahn 1998). One other remarkable alternative to AIPW, that also hits the efficiency bound, is the Targeted Maximum-Likelihood Estimator (TMLE), which can also be used to adapt ML methods for treatment effects (Van Der Laan and Rubin 2006).

2.1.2 Heterogenous Treatment Effects

Ultimately, up until now we made the assumption that there is only one treatment effect for everybody in the sample. That is the reason we have constant ATE in our expressions. As it is mostly unrealistic to think everyone has benefited equally from the

intervention, I introduce heterogenous treatment effects in what follows. In fact, the main premise of this thesis is the identification of the heterogenous treatment effects. I begin by slightly modifying the expression for the ATE:

$$Y_i = \mu(0, X_i) + \tau_i W_i + \epsilon_i$$
$$\tau_i = \tau(x) + e_i$$

Even after ensuring that all our assumptions above hold, the identification of the τ_i requires some technical conditions including independence in higher moments from the error term. The best approach is to find subgroups of individuals for whom the effect is constant, rather than to grapple with with individual treatment effects. These subgroups of individuals are usually defined by some common values of their characteristics X_i, which also adds to the interpretation of the whole heterogeneity story. It is more relevant to discuss the effect of a marketing campaign for the regular customers of an younger age, since it allows us to target customers systematically.
Hence, this is formulated as the Conditional (on characteristics) ATE (CATE):

$$\tau(x) = E[Y_i^1 - Y_i^0 | X_i] =$$
$$= E[Y_i^1 | X_i = x, W = 1] - E[Y_i^0 | X_i = x, W = 0] =$$
$$= \mu(1, x) - \mu(0, x)$$

The key to recovering the treatment effects $\tau(x)$ under heterogeneity is to proceed with a rich enough model specification to capture all the possible subgroups along which the effect can be constant. This can be particularly problematic with linear specification, yet if one has a reason to believe the outcome equation is indeed linear, it is imperative to saturate it with interactions among all the variables to reduce the CATE bias to the minimum. The main task is to identify those subgroups where the treatment effect is relatively stable for every unit, yet variation in covariates X_i is adequate to arrive at that subgroup. When running a linear model without all the possible interactions, one fails to identify those subgroups, and hence the $\hat{\tau}(x)$ remains biased.

This bias is different from the omitted variable kind or confounding bias. Even with access to the full set of X_i that are relevant to guarantee unconfoundedness, the bias persists unless one minimizes the variance of the error term e_i. That amounts to saturating the functional form of $\tau(x)$ such that it maximizes the explained variance by

it. The reason variance is important is that it naturally translates into heterogeneity:

$$Var(\tau_i) = Var(\tau(X_i)) + Var(e_i)$$

As a result, given that the covariance between the two terms on the right-hand side is zero (unconfoundedness), ideally a good approximation of τ_i can be achieved by maximizing the variance $Var(\tau(X_i))$. This conclusion is essential to accomodate any machine learning algorithm into the estimation of heterogenous treatment effects. I return to this expression later, after introducing a brief background on machine learning (ML).

2.2 Background on Machine Learning

In contrast to most of traditional econometrics, machine learning (ML) usually does not rely on the assumption that the sample data follows a particular distribution (Athey and Imbens 2019). Rather, the ML models are data-driven, which is just another word for "non-parametric". Economics community has been rather slow to adopt ML methods, partially due to cultural reasons set out by the top publishing journals (Athey and Imbens 2019). As such, methods without attractive large-sample properties have not been used in the current research, even though ML offers superior performance in complex prediction tasks (Athey and Imbens 2019; Mullainathan and Spiess 2017; Athey 2018).

However, ML has been gaining popularity over the last decade among economists. The main advantage of ML is, certainly, the superior predictive power, which in part can be attributed to the performance evaluation criteria in ML. Compared to traditional econometric prediction, ML methods follow the objective of minimizing the error from the next observation, which the model has not seen yet. The final model is evaluated on a test sample, i.e. the data that the model has not dealt with so far. It allows the practitioner to notice the high variance and overfitting accumulated in the training stage (Athey and Imbens 2019; Mullainathan and Spiess 2017).

The evaluation processes in ML allow for a systematic selection of the most suitable models for a given task. In contrast, traditional econometrics research typically focuses on specifying a single model to estimate a particular parameter of interest, with

confidence intervals derived from asymptotic theory. However, these estimates can be highly sensitive to the chosen model specification, necessitating robustness checks by researchers. A growing concern arises from the potential disincentive to report controversial robustness checks that might invalidate the initial estimates (Athey 2018).

Herein lies an additional advantage of adopting ML methods: when employed properly, they act as a safeguard against distortions caused by **data mining and p-hacking**. The terms in bold in the previous sentence have been generally used as a pejorative, however, nowadays the attitudes toward them are changing in part due to the ML solutions that minimize human interference and further manipulation of results. By subjecting models to careful evaluation and validation, ML methodologies can help mitigate the risks of cherry-picking favorable results and enhance the credibility of research findings. Thus, ML methods can serve as a valuable tool to promote more robust and transparent practices in empirical research, guarding against spurious conclusions driven by questionable model specifications (Athey 2018; Athey and Imbens 2019).

The supervised ML models are well suited for viewing the relationship between dependent variable and many predictors. In particular, regularized regression approaches allow for a faster and efficient prediction than OLS. When objective is to predict the next unit, OLS does not have attractive properties given the number of variables is greater than three, let alone when it gets closer to the sample size.

In the "big-data" settings, which is increasingly more common nowadays, the number covariates K approaches or exceeds the sample size N, rendering estimation of any one parameter in the model almost infeasible with traditional methods. In semi-parametric econometrics, kernels have been a dominating approach. However, as the number of covariates reach and/ or exceed 20, the kernel methods become computationally tedious and inaccurate under the curse of dimensionality (Athey and Imbens 2019). One option is to adopt the regularized regression models, such as LASSO or Ridge to explicitly penalize the inclusion of the additional variables in the model. With a LASSO, for instance, penalty drives some parameters to a zero. While the regularized regression is similar to OLS in the sense that it is suitable for linear and low-dimensional CEFs, with high-dimensional and non-linear applications, the ML literature can offer unsupervised learning methods for dimension reduction. These cases can be better

handled by clustering techniques such as K-means clustering or Principal Component Analysis (Athey and Imbens 2019; Mullainathan and Spiess 2017). In general, almost all ML methods are well-adept to curse of dimensionality, but for the purposes of this thesis I focus on tree-based methods mainly.

2.2.1 Decision Trees

The ML literature proposes tree-based algorithms as an alternative to kernels. These **Classification and Regression Trees** (CART) by Breiman et al. (1984) operate by splitting the covariate space via optimizing the objective function. In contrast to kernels, a tree optimizes by splitting along one variable at a time, and to arrive at that one optimal split, it evaluates the change in objective function at each value of every single variable. Once the optimal value and the variable is found, the algorithm proceed with next split, same as before, but now for both subsets of the initial split. The process continues until there is no significant change in objective function achieved with a split. For a regression task, a tree minimizes MSE, choosing a value x of X to split on:

$$\mu_L(x) = E\ [\ Y_i\ |X \in leaf(X_i \leq x)\ ]$$
$$\mu_R(x) = E\ [\ Y_i\ |X \in leaf(X_i > x)\ ]$$

and the split naturally maximizes the weighted difference

$$\mu_L(x) - \mu_R(x)$$

because that is when the MSE is minimized:

$$MSE = E[(Y_i - \hat{Y})^2] = E[((Y_i - \mu_L(x))^2] + E[((Y_i - \mu_R(x))^2]$$

Advantages of using the trees include but are not limited to the ease of interpretation, better resistance to the curse of dimensionality, and capacity to handle categorical variables without prior transformation. The last point also adds to the fact that it is an extremely simple to implement as an off-the-shelf algorithm, and does not require rescaling and standardization *apriori*. Furthermore, in contrast with kernel methods, decision trees naturally allow automatic variable interaction. As a result, the model

significantly adds to the accuracy of a prediction (Athey and Imbens 2019).

Despite the advantages in ease of use and readability, the CART models tend to have a large variance, are unstable, and rather discontinuous. This directly stems from the fact that it overfits the training sample. To see why the model is prone to overfit, one can start by decomposing the Mean Squared Error (MSE) objective into bias and variance:

$$MSE = E[(Y_i - \hat{Y})^2] = Var(\hat{Y}) + Bias(\hat{Y})^2$$

While $\hat{Y}$ is estimated on a training sample, the evaluated MSE will be minimal, since the model maximizes variance explained and minimizes the bias. Allowing the model to reach such level of complexity that perfectly explains the training data is not well-tailored for the new data that have not been observed yet. Thus, one should control for the model complexity with a penalty, and separate the training data from validation/testing sample on which the model is evaluated later.

Most widely used penalization for model complexity is adding a norm of parameter values to the MSE, as in ElasticNet (of which Ridge and LASSO are special cases when norm is 2 and 1 respectively). To fend the overfitting in trees in particular, pruning can be used to get rid of the nodes that add little value to the optimization goal. Another method is to implement Cross-Validation (CV), i.e. split the training sample into subsamples and train the algorithm on each, depending on the type of CV.

The more recent surge in the literature is to use cross-fitting to fend against overfitting. It emphasizes splitting the data into several folds, leaving one out and training the model on the rest. That model is then used to make predictions for the left-out fold, and this procedure repeats for all the folds. Thus, predictions made for observation i, come from the model trained on out-of-sample observations. This is critical when the number of covariates is close to the sample size, and one observation can have strong distortionary impact on the final prediction for variable X_i (Athey and Imbens 2019; Chernozhukov et al. 2018; Belloni, Chernozhukov, and Hansen 2014b).

2.2.2 Ensemble Methods and Random Forests

Although I characterized single models so far, the most noteworthy results in prediction tasks are attributable to ensembles of several models (Athey and Imbens 2019). These ensemble methods are extremely popular among practitioners as one can average across different models to gain incremental improvements in the objective function. One of the most famous, easy-to-implement ensemble methods is the Random Forests of Breiman (2001). Other methods include bagging, bragging, **subsampling**, stacking and boosting (see Hastie et al. (2009) for more in-depth discussion of those). For the purposes of this thesis, I focus on Random Forests alone.

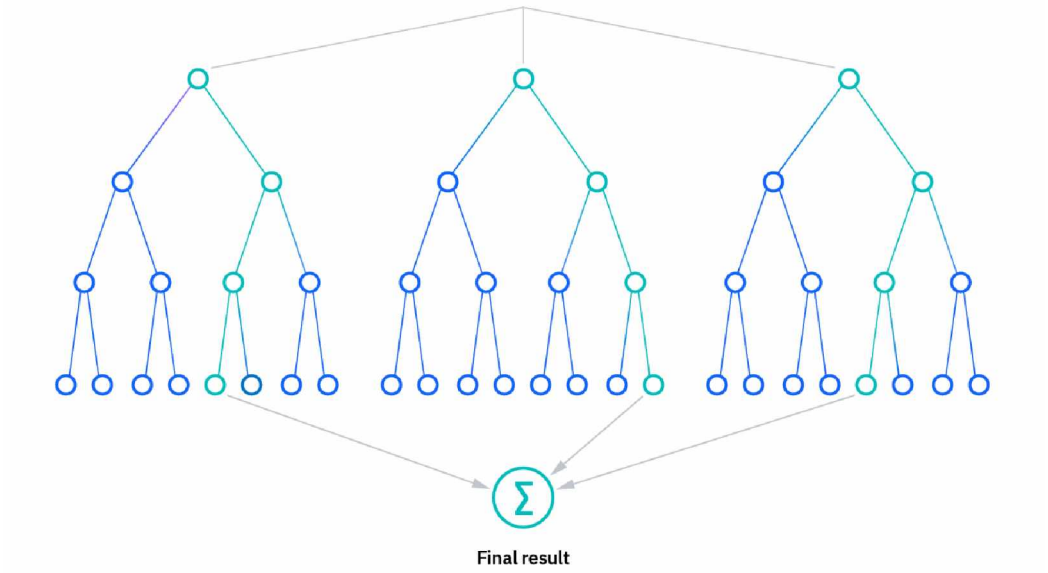

Figure 1: What is Random Forest? Source: IBM

Random forest is an extension of the CART, also improving on the variability of the prediction of a tree (Breiman 2001). The idea is to **bootstrap-aggregate** a lot of trees, hence the forest, and minimize the variance, while keeping the bias constant. It also **introduces smoothness** over discontinuity of a tree. One distinct feature of this ensemble method is the fact that number of covariates, K , is fixed, and the algorithm randomly shuffles them from one tree to another. This helps decorrelate trees and make them seek different specifications to model the outcome. As a result, we get averages of thousands of individual trees, which individually are prone to overfit and may even be inaccurate, but as an ensemble result in a robust prediction machine (Breiman 2001).

2.2.3 Predictive Policy Problems

Based on the preceding discussion, it is pertinent to highlight that certain policy-relevant questions could be rephrased as prediction tasks and directly estimated using ML algorithms. For instance, Kleinberg et al. (2015) illustrate the motivation for using the off-the-shelf ML for prediction in policy settings. As an empirical example, they present the decision of Medicare to cover for joint-replacement surgery for particular patients. This problem requires knowing the patient's chance of survival after the surgery. Then, at a given cost of surgery and estimated benefit of it, the decision reduces to comparing the two under the probability to survive. Intuitively, if a person is already at high risk of death due to other underlying conditions, the potential benefit of providing surgery for that individual may be outweighed by the associated risks and costs. In such cases, medical professionals and policymakers must carefully consider the overall health status of the person, the potential impact of the surgery, and the likelihood of a successful outcome. It becomes essential to assess the risk-benefit ratio to make informed decisions that prioritize the well-being of the individual and maximize the efficient use of medical resource (Kleinberg et al. 2015).

The authors analyzed a sample of (N~100,000) patients who had a joint-replacement surgery covered by Medicaid, and used LASSO with more than 3,300 variables (risk factors, demographics, health and healthcare history, and etc.) to predict the risk of not surviving within a year of the surgery. By replacing the riskiest 10% of the patients with median-risk individuals, Kleinberg et al. (2015) show that the predicted risk scores can save nearly $160 million per year, by postponing the surgery for high-risks who would not have died immediately (Kleinberg et al. 2015).

Another example of the direct application of ML predictions is demonstrated in the study conducted by Glaeser et al. (2016) on restaurant inspections. In this research, the authors organized a competition, fostering a collaborative atmosphere akin to the spirit of the ML community, to outsource the prediction algorithm. Participants received a dataset comprising more than 30,000 inspection cases and customer review data from Yelp for restaurants in Boston.

The winning algorithm, a combination of random forests and boosted trees in an ensemble approach, proved highly effective in targeting restaurants for inspections.

It demonstrated the potential to enhance inspection productivity by up to 50%, by prioritizing those restaurants with a higher likelihood of health or safety violations (Glaeser et al. 2016).

The examples presented on optimal resource allocation only touch upon one aspect of a larger and more complex issue. When examining targeting rules under cost constraints, it is crucial to consider the heterogeneity of the intervention's effects. In cases like hygiene inspections, the focus should be on the impact of the new inspection policy and how it varies among different establishments or units (Athey 2017).

Policy rules generated by algorithms may be stable only if the underlying units demonstrate consistent behavior over time. However, it is vital to recognize that these rules typically pertain to predicting the CEF $\mu(1, X)$, overlooking the counterfactual $\mu(0, X)$.

To establish proper targeting rules, it is essential to incorporate both the control and treatment groups and jointly estimate the model, thereby addressing selection issues and potential confounding factors.

Blake, Nosko, and Tadelis (2015) provide a critical exposition of thinking of policy evaluation as a prediction task in online advertising industry, another domain where policy targeting plays a crucial role. Ebay's implementation of paid advertising in search engines like Google and MSN's Bing initially suggested an impressive 1600% return for each dollar spent, based on predictive estimates correlating clicks with sales (Blake, Nosko, and Tadelis 2015; Athey 2017).

However, to examine the robustness of these estimates and understand whether the ads truly influenced consumer behavior, Ebay conducted an experiment. They stopped the ads on one platform while using the other as a control. The results revealed that customers who spent money after clicking on the paid ads would have spent the money anyway because they were already aware of the brand. Consequently, the causal effect of the paid ads was found to be -63% according to Blake, Nosko, and Tadelis (2015), indicating negative returns to advertised search.

The study by Blake, Nosko, and Tadelis (2015) exemplifies several important themes, including the reduced costs of online experimentation, the involvement of big tech firms in econometric questions, the risks of improper methods, and the value of credible policy evaluation. Furthermore, in an interview with Scott Cunningham, one of the

authors, Steven Tadelis, highlights the changing landscape in the industry due to an increasing number of economists and the adoption of more credible methods to address issues such as selection bias, endogeneity, and spillovers for inference (Cunningham 2022a). In her interview, Susan Athey highlights the natural questions that arise from large-scale experimentation in large private corporations. She emphasizes the significance of adopting design-based and structural approaches, effectively integrating them with methods commonly used by computer scientists (Cunningham 2022b). Indeed, with the increasing availability of big data and the integration of credible and rigorous economic foundations, the development of methods and designs for conducting research in big-data settings is a natural progression. Hence, in the section below I expand on how the ML methods have been properly incorporated into the policy evaluation questions.

2.3 Causal Machine Learning

Clearly, ML methods have several potential advantages in prediction tasks over traditional econometric practices. They can help mitigate overfitting, offer wide range of algorithms with better performance, and allow for estimation when the number of covariates K is close or exceeds the sample size N. Yet when trying to adopt ML into causal questions, we need to be explicit with our goals. When the number of covariates is far greater than the sample size, one option is employ regularization. Regularized regressions such as LASSO or ElasticNet would work quite well. Alternatively, one can employ unsupervised learning to cluster variables, reducing the dimensionality for further analysis. Both these options allow us to reduce the number of dimensions for further use. The key idea is that ML methods are workable both for **post-selection inference** and **prediction**. If we are concerned with estimating the ATE and have access to relevant variables to account for potential unobservable confounders influencing the outcomes, it is feasible to address both objectives simultaneously.

2.3.1 Post-LASSO Selection

Regarding LASSO/Ridge estimators, their consistency is assured under the assumption of **approximate sparsity**, and some extra technical conditions. This notion intuitively

suggests that only a fraction K_s of the variables is actually significant, conveying substantial information. Consequently, two approaches can be adopted:

a) Directly estimating all variables employing LASSO, which is a straightforward albeit somewhat naive method.

b) Utilizing LASSO as a selection method, which proves to be more engaging. This method permits the identification of the subset K_s containing the variables with critical information, subsequently enabling the implementation of Ordinary Least Squares (OLS) regression specifically on this selected subset (Belloni, Chernozhukov, and Hansen 2014b).

Post-LASSO literature is heavily focused on adapting the penalized estimators for the estimation of low-dimensional ATE when the researcher is in a setting with large number of covariates, or is highly uncertain about the functional form in which those covariates should be included. Belloni, Chernozhukov, and Hansen (2014a) demonstrate how highdimensional inference can be done properly. One of their applications is revisiting the Donohue and Levitt (2001) study on the effect of abortion legalization on crime rates. Donohue and Levitt (2001) used differences-in-differences (DiD) specification with state and time fixed effects. In their setting, the estimate for the causal effect rely on the assumption that the time-variant state characteristics capture all the remaining variance, hence leaving clear causal effect of the intervention (Belloni, Chernozhukov, and Hansen 2014a). They implement LASSO to select variables with non-zero coefficients and then conduct OLS regression with selected variables. The results show much tighter confidence intervals compared to the "kitchen sink" OLS with 284 variables, but they also highlight increased uncertainty surrounding the causal effect of abortion legalization (Belloni, Chernozhukov, and Hansen 2014a).

Overall, Belloni, Chernozhukov, and Hansen (2014a) showcase the potential of using Post-LASSO techniques to refine causal inference in settings with a large number of covariates, providing more precise estimates and addressing concerns related to overfitting.

Another noteworthy application of the (double) post-LASSO is Bach, Chernozhukov, and Spindler (2018), in which authors try to capture the heterogeneity in the gender wage gap. Bach, Chernozhukov, and Spindler (2018) tried to model the wages using a high-dimensional regression of socioeconomic factors, including marital status, region,

education, tenure, occupation and industry, religion, race, children, and many more. Their data is a 1% representative sample of the US population provided in the American Community Survey (ACS). In scenarios with a large number of covariates, evaluating heterogeneity through two-way interactions can be computationally expensive. To address this, the researchers implemented the double-LASSO post-selection method introduced by Belloni, Chernozhukov, and Hansen (2014b). This approach proved more useful than a simplistic mean-decomposition of the gender gap, as it allowed for a detailed analysis of which subgroups of women face the most significant challenges.

Per their findings, the gender gap in classic human capital variables like education and experience was relatively small. However, significant variations were observed across different industries, occupations, and family compositions. Notably, women in financial and public service sector jobs were found to be more underpaid compared to men with similar observable characteristics (Bach, Chernozhukov, and Spindler 2018).

This example holds particular importance for the purpose of my thesis, and I will revisit it in the application section to further explore its implications.

2.3.2 Double/debiased ML

The problem is that LASSO/Ridge are still parametric, that is, model specific. They heavily depend on the functional form assumptions. To estimate finitely many interactions and higher order relationships among those K variables in a data-driven manner, tree-based methods or neural networks are better suited.

An alternative way to filter noise is to use a control function that incorporates all relevant variables simultaneously. This comprehensive approach aims to account for potential confounding factors and improve the accuracy of the analysis. Instead of obtaining a subset K_s for subsequent OLS, we adopt a different approach by partialing out the intricate correlations from the variables in Y (outcome) and W (treatment), which yields biased post-ML error terms. The key to debiasing lies in the Neyman-orthogonal score, which helps remove the effects of nuisance predictions, and turns to a residual-on-residual analysis to achieve unbiased estimation. This is a more generic definition that allows us to construct orthogonal score for any estimator (Chernozhukov et al. 2018). One can do so by constructing a loss function with a regularizer which would be first-order insensitive to the nuisance parameters. The AIPW we discussed

before is one of the well-known orthogonolized scores. This, along with the cross-fitting we discussed previously, under unconfoundedness allow for the debiased estimate of ATE given we have $K >> N$ (Chernozhukov et al. 2018).

As a result, given the doubly-robust score τ_{IPW}, the estimate $\hat{\tau}_{IPW}$ converges to the oracle value, if one of the nuisance parameters $e(X_i)$, $\mu(W_i, X_i)$ converges at a faster rate $n^{1/4}$. One can estimate those nuisance parameters with any conventional ML algorithm, yet it is important to ensure there are no inconsistencies coming from the choice of the method. As such, using cross-fitting for those parameter accounts for the idiosyncrasies. To avoid confusion, I define cross-fitted parameter predictions $e^{-i}(X), \mu^{-i}(W, X)$ for a unit i in fold K of our sample, where the model is trained on all the observations outside this unit's fold.

A noteworthy example of double/debiased ML is the study by Farrell, Liang, and Misra (2021). In this research, the authors assess the performance of eight different feed-forward neural networks for estimating the uplift (marketing terminology for the ATE), in the context of an email catalog campaign.

Beyond obtaining valid estimates for the uplift, Farrell, Liang, and Misra (2021) also demonstrate the proper execution of policy targeting using the well-approximated underlying functions. Additionally, their findings provide fast and optimal convergence rates for the neural networks, thus encouraging this approach for causal inference in complex settings (Farrell, Liang, and Misra 2021). This work highlights the potential of neural networks for addressing policy evaluation questions and leveraging double/debiased ML for more accurate and robust causal estimations (Farrell, Liang, and Misra 2021).

2.4 Uncovering Heterogeneity with Causal Forests

In the pursuit of causality, another approach to implementing ML involves modifying the objective function of CART to uncover heterogeneity in the estimates. This thesis focuses on this particular method, which will be described in detail below.

To make trees useful in the potential outcomes framework, it is essential to be clear about the objectives. Traditional trees partition the covariate space by optimizing the objective function, often using MSE for regression tasks, aiming to maximize prediction accuracy. However, in the context of causal inference, prediction is not always the

primary goal, and adjustments are necessary. Furthermore, the lack of access to the ground truth hinders the evaluation of the model on a test set. I start by rewriting the expression from Section 2.1.1 :

$$Var(\tau_i) = Var(\hat{\tau}(X_i)) + Var(e_i)$$
$$\hat{\tau}(X_i) = \mu(1, x) - \mu(0, x)$$

The baseline method for partitioning the covariate space would be the kernels as **K-NN**, which calculate the CATE as

$$\hat{\tau} = \frac{1}{k} \sum_{S_1(x)} Y_i - \frac{1}{k} \sum_{S_0(x)} Y_i$$

where $S_1(x)$ is the set of k nearest treated units to the given x. This estimator is unbiased and has attractive asymptotics, but as mentioned already, fails as the covariate space expands in dimension.

Random Forests, on the other hand, are excellent at partitioning, yet often used without any regard to their asymptotics. Without the latter, it is hard to infer a relationship and construct confidence intervals. In the potential outcomes framework, where we cannot observe both states for an individual and hence test the accuracy of the algorithm, it is imperative to refer to asymptotics to conduct hypothesis testing. Wager and Athey (2018) provide first formal results on asymptotics of Random Forest estimates.

Furthermore, several ML meta-learners have been developed that also start with targeting estimation of the CATE through CEF $\mu(w, x)$. **S-learner**, for example, estimates a single model for $\mu(W, X) = E[Y|X, W]$, while **T-learner** estimates separately $\mu(0, x),\ and\ \mu(1, x)$ for treatment and control groups. Both meta-algorithms then take the difference for CATE $\hat{\tau}(X_i) = \mu(1, x) - \mu(0, x)$.

It is apparent that, these models fail to recognize the nature of the data, mainly lack of the account of the selection bias. T-learner fails by fitting each of the CEFs separately, which only holds under simple CATE is the difference. S-learner behaves more optimally by fitting one CEF for both groups, but performs well when the groups are balanced. Künzel et al. (2019) propose the **X-learner**, that on the other hand,

constructs the counterfactual outcome for each unit using predictions from separately estimated $\mu(0,x),\ and\ \mu(1,x)$. Those counterfactuals are subtracted from the actual outcome value for each unit , and the CATE is taken as the weighted difference $\hat{\tau}(X) = g(x)\hat{\tau}_1(X) - (1-g(x))\hat{\tau}_0(X)$.

Künzel et al. (2019) reanalyze the effect of mail-induced peer-pressure on the voter turnout in the US. The original authors find no evidence of heterogeneity in the effects, whereas the program turned to be quite effective raising turnout by 8.1%. Capturing heterogeneity in the sample properly would allow for more targeted mailing for the next campaigns (Künzel et al. 2019). Applying the three meta-learners (S, T, and X) Künzel et al. (2019) conclude that for the group of people who voted 3 times in the past 5 elections the impact of the social pressure is the highest. They uncover nearly 50% more turnout due to the mailing among those households (Künzel et al. 2019).

While X-learner performs quite well under randomized treatment, it fails to construct counterfactuals from X_i in observational studies (Künzel et al. 2019). Causal Trees help recover from this issue by taking the $g(X)$ as the propensity score, thus estimating CATE in a doubly-robust fashion.

Below I describe how to grow Causal trees into a forest, both initial versions of Causal Forests and the later generalization of them into Generalized Random Forests by Athey, Tibshirani, and Wager (2019).

2.4.1 The Mechanics of a Causal Tree

Athey and Imbens (2016) show how to modify the objective properly and introduce the Causal Trees. Essentially, the idea is to feed the CATE as an objective function to the tree, so that it maximizes the heterogeneity with each split.
Let us start by defining average outcome

$$\mu(w,x;\Pi) = E\ [\ Y_i(w)\ |X \in l(x;\Pi)\ ]$$

where Π denotes the tree structure, and $l(x;\Pi)$ the leaf of the tree where the given x falls.

$$\tau(x;\Pi) = E\ [\ Y_i(1) - Y_i(0)\ |X \in l(x;\Pi)\ ] = \mu(1,x;\Pi) - \mu(0,x;\Pi)$$

then is the average treatment effect, conditional on X.

The estimated analogs, $\hat{\mu}$ and $\hat{\tau}$, are the **mean outcome within the leaf of the tree for treatment/control**, and the **difference in means between the groups in the same leaf**, respectively. This was the definition for the initial Causal Tree algorithm proposed by Athey and Imbens (2016). Later on, they incorporated the doubly-robust AIPW score for the treatment effects as already mentioned:

$$\tau_{AIPW}(X_i) = E\left[\frac{W_i\ (Y_i - \mu^{-i}(1,X))}{e^{-i}(X)} - \frac{(1-W_i)\ (Y_i - \mu^{-i}(0,X))}{1-e^{-i}(X)}\right] \\ -E[\ \mu^{-i}(1,X) - \mu^{-i}(0,X)\ ]$$

Thus, the counterfactuals are constructed via inverse-propensity score weighting and CEFs, which also account for the selection bias, but requires the common support. Drawing on the results of Chernozhukov et al. (2018), if one of the nuisance parameters $e^{-i}(X)$, $\mu^{-i}(W,X)$ converge at a rate $n^{1/4}$, the whole score is unbiased and converges to the actual treatment effect value. This encourages the use of ML algorithms with the greatest predictive power in a setting as an intermediate step, out-of-sample predictions of which are then plugged into the final model.

Next, define MSE as

$$MSE_\tau(S^{te}, S^{est}; \Pi) = \frac{1}{|S^{te}|} \sum_{i \in S^{te}} \Big(\tau_i - \hat{\tau}(X_i, S^{est}; \Pi)\Big)^2 - \tau_i^2$$

S^{te}, S^{est} are the testing and estimation samples respectively. The difference here with the traditional train/test split in modeling is an aspect called **honesty**. A tree is honest when none of the observations used for the training, i.e. growing it appear in the validation sample. This is the second criteria, apart from MSE modification that allows to draw asymptotic conclusions in the CF framework.

Additionally, the honest approach helps with handling the outliers. As such, in first step, the tree splits on the outlier value of Y and maximizes heterogeneity in treatment effect. Yet, if we were to take the treatment effect in the leaf built around outliers, it would yield a spurious and large (in absolute value) numbers compared to the average effects. Whereas when we simply take the tree structure from the first step, and use it to actually estimate the effects the bias diminishes. Since it is a novel data sample and

does not contain the same outliers.

Note that by taking the *MSE of the treatment effect*, not the custom one over the outcomes, we account for the fact that heterogeneity in outcomes and heterogeneity in treatment effects are caused by different sets of variables. As such, a person might be assigned a new drug by the physician based on the risk factors and how far along the disease is progressed. Yet to explore the variation in how that drug worked it serves a richer insight to check for the patient's environment, diet, personal relationships and etc. Although those factors are rarely properly accounted for, heterogeneity along them could reveal a better treatment assignment mechanism for the next trials. Hence, heterogeneity provides value of exploring the new features for policy learning.

Moving on with MSE_τ, it appears that the expression contains an infeasible term τ_i. However, the tree algorithm comes in handy, since τ_i is stable within the leaf. That is, the objective function involves a constant term, and hence subtracting the τ_i^2 in the above expression won't change anything from the optimization viewpoint. As a result, our objective function is

$$EMSE_\tau = E_{test}\left[E_{train}[\ MSE_\tau]\right]$$

and using the fact that

$$E_{test}[\tau_i] = E_{test}[\hat{\tau}]$$

we simplify the MSE estimator to

$$\begin{gathered}\hat{MSE}_\tau(S^{te}, S^{est}; \Pi) = \frac{1}{|S^{te}|}\sum_{i\in S^{te}}\left(\tau_i - \hat{\tau}(X_i, S^{est}; \Pi)\right)^2 - \tau_i^2 = \\ = -\frac{2}{|S^{te}|}\sum_{i\in S^{te}}\hat{\tau}(X_i, S^{est}; \Pi) * \hat{\tau}(X_i, S^{est}; \Pi) + \frac{1}{|S^{te}|}\sum_{i\in S^{te}}\hat{\tau}^2(X_i, S^{est}; \Pi) = \\ = -\frac{1}{|S^{te}|}\sum_{i\in S^{te}}\hat{\tau}^2(X_i, S^{est}; \Pi)\end{gathered}$$

Consequently, when making each split, the algorithm looks for the value of the X_i that maximizes the heterogeneity of the treatment effect. In contrast with all the previous methods, this objective function optimizes over a higher order moment of the CATE expression, directly maximizing the variance and finding subgroups via recursive partitioning, where the treatment effect is relatively stable for everyone.

2.4.2 Types of Trees and Growing them into a Forest

Wager and Athey (2018) present two types of causal trees, the so-called double-sample trees, very well suited for studies with treatment allocation mechanism as good as random. In the first stage, the tree optimizes on the difference between the outcome regression models for two groups $\hat{\tau} = \hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)$.

Wager and Athey (2018) also suggest propensity trees for observational studies, which implies fitting a tree by predicting the treatment dummy, completely ignoring the outcome, and then estimating the $\hat{\tau}$ on the test set:

The process described in the previous section outlines the mechanics for one causal tree. To grow a forest, one needs to bootstrap-aggregate (bagging) many more trees, and decorrelate those imposing limits on the number of variables used for each tree. However, to reach the desired asymptotics, it is recommended to use subsampling instead of bagging, as follows:

1. From the original sample draw a random *subsample* s_b (**without replacement**)

2. Build a tree:

- start by splitting the subsample s_b further into training and test parts
- select the smaller subset m out of all K variables to use for this tree
- choose each split such that it minimizes $M\hat{S}E_\tau = -Var(\hat{\tau}(X_i))$
- continue until no meaningful split is available/ minimum number of units per leaf reached
- estimate $\hat{\tau}$ on the test set and save as $\hat{\tau}_{s_b}$ 3. Repeat the steps above B number of times, growing a forest with the total of B trees and saving B estimates for the $\hat{\tau}$ 4. Take the average of all those estimators $\frac{1}{B}\sum_{b \in B} \hat{\tau}_{s_b}$

A tree might be heavily reliant on data and sample size and hence validation splits can be viewed as wasting the data budget. While a forest due to the its subsampling (or bagging) nature, allows the practitioner to avoid withholding a part of the sample for validation, because each tree works only with a subset of the whole sample and we have out-of-bag sample to validate its estimates. This is a nice perk especially when working with limited data budget, so that an individual's own outcome does not influence its subgroup assignment. As a result, the out-of-bag estimates in the forest can substitute the need for cross-fitting procedure.

The `grf` package in R (Tibshirani et al. 2022) allows implementing the CF for doubly-robust estimation. The nuisance parameters, if not explicitly given, are calculated using the regression forests. However, by nature the algorithm does not estimate the propensity scores per se, it partitions the covariate space in a way that best imitates the outcomes of propensities. That is, we do not get to see $e(X_i)$ and plug it to estimate the τ_i, we get the result with solving a classification task as a workaround.

2.4.3 Main Results

Wager and Athey (2018) derive the asymptotic results for both causal forests and random forests, given that the following conditions hold:

- Honesty
- Subsampling the trees for the forest
- Continuous covariates
- The response function $\mu(X_i)$ is Lipschitz-continuous

Forests are asymptotically Normal and centered (Wager and Athey 2018):

$$\frac{\hat{\mu}_n(X) - \mu(X)}{\sigma_n(X)} \rightarrow_d N(0,1) \; ; \qquad \sigma_n^2(X) \rightarrow_p 0$$

2.4.4 Drawbacks of Causal Forests and GRF

The two methods of growing a CF differ in the way they handle the nuisances. The first method is not well-suited for addressing pure confounding, which can lead to biased estimates of causal effects. On the other hand, the second method initially grows a propensity tree, effectively managing confounding issues. However, it struggles to perform optimally in scenarios with high heterogeneity, potentially compromising the accuracy of its estimates in such cases. Balancing the trade-offs between these methods is crucial to ensure reliable and unbiased estimates of causal effects in different contexts (Athey, Tibshirani, and Wager 2019).

Athey, Tibshirani, and Wager (2019) develop a generalization of tree-based ensemble methods, of which the CF is a specific case. Taking the modifications further, the authors introduce the Generalized Random Forest (GRF) for causal inference. Although operating under slightly different objective function, the GRF algorithm significantly improves upon the shortcomings of the two CF approaches. Furthermore, it performs

well in the setting with both high heterogeneity and pure confounding.

This success is due to local recentering nature of their algorithm. Where in the previous versions we used nuisance functions to estimate $\hat{Y}$ *and* $\hat{W}$ and then plug it into the CF, now we add one more step: take nuisance parameters and estimate the out-of-bag residuals $\hat{Y}_i = Y_i - \mu^{-i}(X)$ and $\hat{W}_i = W_i - e^{-i}(X)$. These residuals are then imputed to train GRF and get the treatment effects. This is the value of orthogonolization proposed by Robinson (1988). To understand it better, let us recall that before, the algorithm had to fit two disjoint CEFs $\mu(0, X)$ for control and $\mu(1, X)$ for treatment groups respectively. This leads to flawed results when the degree of confounding or heterogeneity increases Athey, Tibshirani, and Wager (2019). Robinson (1988) proposed the following modification in a partially linear model:

$$m^{-i}(X) = E[Y_i|X = x] = E[Y_i^0|X] + E[W|X]\tau(X) =$$
$$= \mu^{-i}(0, x) + e^{-i}(x)\tau(x);$$

That is, now the CEF is estimated without conditioning on the treatment status, and hence we need to adjust for the selection into treatment via the term $e^{-i}(x)\tau(x)$, with the PS. Next, we can rewrite the partially linear model as

$$Y_i = \mu^{-i}(0, x) + W_i\tau(x) + \epsilon_i;$$
$$Y_i - m^{-i}(X) = (W_i - e^{-i}(X))\tau(x) + \epsilon_i$$

Subsequently, the treatment effect $\tau(x)$ is estimated by minimizing the following expression:

$$E[\epsilon_i|X_i, W_i] = E[Y_i - m^{-i}(X) - (W_i - e^{-i}(X))\tau(x)|X_i, W_i]$$

This leads to a new expression for the treatment effect, compared to the original CF:

$$\hat{\tau} = \frac{\sum_i \alpha_i(X)\ (W_i - e^{-i}(X))\ (Y_i - m^{-i}(X))}{\sum_i \alpha_i(X)\ (W_i - e^{-i}(X))^2}$$

where $\alpha_i(X)$ is the frequency with which individual i falls into the same leaf as the test point x we have. That expression uses the adaptive kernel weights to average out the

trees in the forest. When the treatment effect is constant, the weight is also constant, and is equivalent to running OLS with orthogonolized values. It is noteworthy how the problem of estimating constant versus heterogenous treatment effects is evident here. If the goal is to estimate the constant effect, one could easily assume equal weights, or focus on the subgroup with the strongest signal. Whereas heterogenous treatment effects are spread out through subgroups, and it is imperative to assign corresponding weights to them depending on the strength of the signal. By the strength of the signal I mean the variance of the treatment effects (heterogeneity). Hence, the problem converges to identifying the subgroups where the the effect is constant within, but is highly variable across them. This is exactly the mechanism behind the trees as previously discussed. The corresponding weights then are the $\alpha_i(X)$.

Moreover, Nie and Wager (2021) develop a quasi-oracle meta-learner for estimating heterogenous treatment effects, the R-learner, allowing to accommodate any other ML algorithm into the structure. It is only an incremental step to employ R-learner, as the `grf` already runs on the similar objective function to their R-loss, it is the moment condition above plus some regularizer (Tibshirani et al. 2022).

However, there are other modifications such as **Local Linear Forests** (LLFs), to correct for the local discontinuities in the original model. Since the CF is a step-function, the predictions on the edges of the covariate support can be poorly modeled.LLFs are designed to exploit the adaptive kernel α_i from the recursive partitioning, and then running a linear model with these corresponding weights (Friedberg et al. 2020). Hence, the LLF can be seen as a more robust alternative to traditional kernel regressions and Friedberg et al. (2020) provide empirical and theoretical benefits of doing so.

As an example of policy learning with Causal Forests, Athey and Wager (2021) demonstrate how one can construct optimal policy rules that are also clearly communicable to the public. Based on the semiparametric theory, given a doubly-robust estimate for the CATE for the whole sample, they showcase how to find policy targeting that minimizes regret. In an empirical setting, they use the data from Greater Avenues for Independence (GAIN) program from 1986 (Athey and Wager 2021). The program served to provide job search training and other educational resources, and had a randomized assignment mechanism (Athey and Wager 2021). Participants' results were recorded for the following 9 years, and evaluations showed significant benefits to the average wages

ex-post (Athey and Wager 2021).

Despite the fact that allocation to GAIN program was random, it differed for each county, and the pooled county analysis would be confounded. Athey and Wager (2021) use 54 variables to assume unconfoudedness and implement the doubly-robust Causal Forest to arrive at the income gains for the whole sample. The doubly-robust specification results in closer estimates to the original treatment effect than naive difference in samples, emphasizing the value-added in controlling for confounders (Athey and Wager 2021). The key takeaway, however, is that they subsample units who would have benefitted more than the whole sample (i.e. CATE greater than the total ATE). An example of their policy rule is in the figure below, which is reported to yield an additional 0.08 benefit to the original ATE of 0.14 for the whole sample (Athey and Wager 2021).

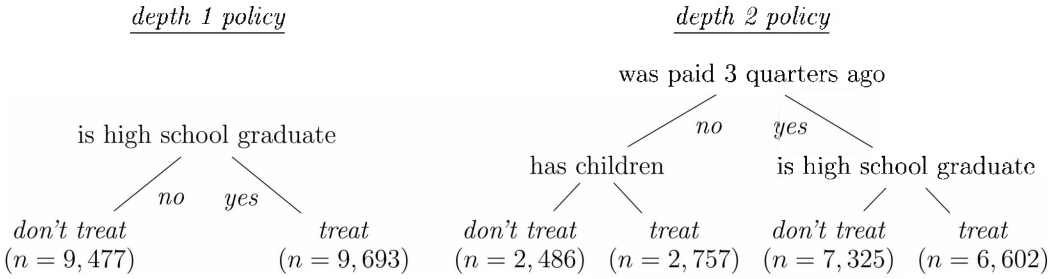

Figure 2: Policy trees for the GAIN study, Source: Athey and Wager (2020)

A recent method developed to address the question of "whom to treat" is the **rank-weighted Average Treatment Effect** (RATE), proposed by Yadlowsky et al. (2021). RATE is a model-agnostic approach that aggregates the predictions of CATE obtained from a diverse set of estimators. These CATE predictions are then ranked based on the potential benefit they offer from the treatment, sorting them from the highest to the lowest. Yadlowsky et al. (2021) provide asymptotic confidence for the reweighing process and demonstrate the effectiveness of their approach through real-world examples, including personalized hypertension treatment and uplift modeling. Because of the nature of the medical data, they use censored version of Causal Forests, among other algorithms, to estimate CATE for the sample of SPRINT and ACCORD studies (Yadlowsky et al. 2021). For the uplift modeling, they use large online trial data provided by Criteo and use usual implementation of Causal Forests. In their study, RATE successfully uncovers significant heterogeneity in the marketing context, while in

the medical case, they do not find substantial evidence of heterogeneity (Yadlowsky et al. 2021). This highlights the versatility and potential of the RATE method in identifying subgroups that could benefit the most from an intervention, making it a valuable tool in various fields of study. I demonstrate the application of RATE as a test for heterogeneity in my empirical application.

Up to this point, the focus has primarily been on applications of causal machine learning with selection on observables. However, the objective functions used in these methods demonstrate a remarkable flexibility to incorporate quasi-experimental approaches such as instrumental variables (IV). An illustrative example of this is presented by Athey, Tibshirani, and Wager (2019), who introduce "Instrumental Forests", built on the same principles of honesty, subsampling, and a modified doubly-robust objective.

In their study, Athey, Tibshirani, and Wager (2019) revisited the argument put forward by Angrist and Evans (1996), which used the sexes of the first two children as an instrument for labor force participation. While Angrist and Evans (1996) concluded with a local ATE, Athey, Tibshirani, and Wager (2019) delved further into exploring the heterogeneity in labor supply among mothers who already had two children of mixed sexes.

To extend their analysis, Athey, Tibshirani, and Wager (2019) introduced heterogeneity modeling based on several covariates, including the mother's age at the birth of her first child, her overall age, education, race, and the father's income. Their findings indicated that the negative labor supply was mainly driven by mothers with lower husband's income. However, they also pointed out that the measure of income itself can be endogenous, potentially compromising the further interpretation of the results. Nevertheless, the application of Instrumental Forests demonstrated the capability to uncover important heterogeneity in the context of labor supply, thus showcasing the power of causal machine learning in the presence of instrumental variables. These extensions and modifications (IV Forests, LLF, RATE) are readily available in an easy-to-implement fashion with the `grf` package in R (Tibshirani et al. 2022).

Overall, in this review, I discussed the growing literature behind the estimation of heterogeneous causal treatment effects with machine learning. I started by laying out the definitions of treatment effects in potential outcomes framework, discussed the selection

bias when treatment assignment is not randomized, and reviewed assumptions to cope with it in observational studies. I also defined treatment effects under heterogeneity, on which all the further sections build. With a brief review of particular ML methods, I tried to outline how they can be adapted for different policy problems. Especially, I focused on the estimation of treatment effects under heterogeneity and unconfoundedness using Causal Forests. In what follows I build on the results clarified in this section to recover the heterogeneity in gender wage gap.

3 Empirical Illustration

In this part I analyze the gender wage gap following Huber and Solovyeva (2020). They compare how sensitive the estimates of different methods are, for the sample from National Longitudinal Survey of Youth (**NLSY**) in the year 2000. Their results are particularly interesting for non-parametric estimation of the gap and its decomposition into explained and unexplained parts. Moreover, their setup and assumptions lead to more credible interpretation of the gender gap estimates. However, instead of focusing attention on one number and assuming the difference between the groups is constant, I relax that by testing heterogeneity using Causal Forest (CF).

Implementing CF allows me to estimate variation in the **direct effect** (gender bias/ discrimination) by the design of the algorithm. Hence, I do not focus on the details of decomposition of the explained and unexplained parts of the gap, because the method I use does not deliver the estimates of the total effect.

Furthermore, I do not engage in selection-into-employment-corrective measurements, and what follows is completely based on men and women who had a job at the time of the survey. Huber and Solovyeva (2020) in their analysis reported IV correction for labor market participation, using number of children as an instrument, however the results were not significant. One can clearly use doubly-robust IV forests for addressing this issue with a better instrument, yet explaining the motivation of people to join labor force is beyond the scope of this thesis. I aim to utilize the CF to showcase how one can effectively leverage machine learning for policy learning purposes under heterogeneity. Uncovering heterogeneity allows me to realize the parts of the working population where the discrimination is the highest, that then can be targeted to efficient policy rule, or

scrutinized further in reducing the gap.

The gender wage gap has been receiving growing attention of the community well-beyond labor economists recently (Blau and Kahn 2017). Initial attempts focused heavily on trying to obtain one number for the average gap. Those attempts mostly included few handpicked variables from human capital theory as controls, yet the main attention had been directed at the average difference between men and women (Bach, Chernozhukov, and Spindler 2018). Classic approach to decompose the gender wage gap would incorporate a linear model which is then broken down in the spirit of **Oaxaca-Blinder**, identifying the part of the variance in the gap due to the model specification, and the unexplained part. The latter is usually referred to as the discrimination (Blau and Kahn 2017). Later attempts were targeted at developing more non-parametric framework for decomposition, and DiNardo, Fortin, and Lemieux (1995) is most notable example of this trend. DiNardo, Fortin, and Lemieux (1995) develop a kernel-based semiparametric reweighing estimator for decomposing the density of inequality. Their attempt incorporates ideas of counterfactual densities, bringing more clarity over the Oaxaca-Blinder sample difference, and relaxes the linearity assumption (DiNardo, Fortin, and Lemieux 1995). They also address the assumption that the gender gap is constant for everybody in the sample, by trying to locate the most expressive areas in the density of income (DiNardo, Fortin, and Lemieux 1995).

One can see the similarity with the propensity score weighing, however, estimation of counterfactuals in DiNardo, Fortin, and Lemieux (1995) is not based on credible assumptions, and hence prone to mistaking the two distributions (Yamaguchi 2015). The IPW methods based on a causal identification formulated in potential outcomes or DAG frameworks are better at addressing this shortcoming (Yamaguchi 2015; Huber and Solovyeva 2020).

Furthermore, it has been clear by 2010s that exploring heterogeneity along the covariates is necessary (Blau and Kahn 2017). However, to this date, only few studies address the decomposition of the gender wage gap under heterogeneity. They mostly focus on the one specific characteristic and report the variation in the gap along those (race, ethnicity, occupation, married/not). Thus, they effectively ignore the fact that the difference in wages between men and women can vary simultaneously with many other covariates (Bach, Chernozhukov, and Spindler 2018).

However, analyzing a complex socio-economic phenomenon such as gender wage gap requires a proper formulation. Given the recent surge in the methods allowing to address the estimation when the number of parameters to estimate grows close to the sample size, or even exceeds it (see Belloni, Chernozhukov, and Hansen (2014b), Bach, Chernozhukov, and Spindler (2018)), it is only natural to expect the literature on heterogeneity in the gap to expand.

One noteworthy attempt to capture the heterogeneity in the gender wage gap is Bach, Chernozhukov, and Spindler (2018). The authors tried to model the wages using a high-dimensional regression of socioeconomic factors, including marital status, region, education, tenure, occupation and industry, religion, race, children and more. As the number of covariates becomes large, running two-way interactions of those and assess heterogeneity among them becomes too costly. Hence Bach, Chernozhukov, and Spindler (2018) implement the double-LASSO post selection method by Belloni, Chernozhukov, and Hansen (2014b). Their data is a 1% representative sample of the US population provided in the American Community Survey (ACS). Compared to Oaxaca-Blinder estimates of 14% pay gap, their analysis shows that only a small fraction of women experience the gap of that magnitude. They recover heterogeneity among women with lower education, where median wage gap was at least 29%. Moreover, for married women the difference is 9% to 12% larger than to single women. The gap is also reported to be more severe in finance and professional services industry (Bach, Chernozhukov, and Spindler 2018). Overall, Bach, Chernozhukov, and Spindler (2018) showcase how to locate the severity of the gender wage gap leveraging a rich set of individual level covariates. My aim in this section is similar, but while accounting for heterogeneity is a significant leap forward in the literature, it is imperative to recognize that the issue of identification needs to be addressed. This requires a clear framework to think about the factors affecting and affected by the gender image in the society. I follow the framework formulated in Huber and Solovyeva (2020) to achieve clarity in interpreting my estimates.

3.1 Data

I obtained the data used in the analysis by Huber and Solovyeva (2020) from the **Harvard Dataverse** (Huber 2019), which provides open source datasets for re-analysis purposes. All the variables are indicated for the year 2000, unless explicitly stated otherwise. In their analysis, Huber and Solovyeva (2020) operate with the following set of variables from NLSY 1979 data: The outcome variable of interest (Y) is the **log average hourly wage** in the past calendar year reported in 2000. The set of post-group characteristics X, that potentially mediate the effect of gender on wages, includes marital status, years in marriage, region of residence, urban area indicators, education level, employment history, and health-related factors. Huber and Solovyeva (2020) also use higher-order and interaction terms to reach a more flexible propensity score specification. A more detailed list is given in the Table 1, along with the two-sample t-stats of these covariates for both sexes.

Table 1: Unlogged outcome and covariate means for both sexes, Source: Author's replication based on Huber and Solovyeva (2020)

Variables	Men	Women	p-value of the difference
wage	19.37	14.16	0.00
married	0.57	0.57	0.89
yrs married	6.43	7.54	0.00
North East	0.15	0.16	0.85
North central	0.24	0.24	0.61
West	0.21	0.19	0.25
yrs in current region	14.84	15.25	0.00
SMSA	0.81	0.82	0.57
yrs in current SMSA	13.48	14.20	0.00
Degree highschool	0.46	0.42	0.00
some college	0.21	0.27	0.00
college or more	0.20	0.21	0.41
first job before 1975	0.06	0.05	0.00

Variables	Men	Women	p-value of the difference
first job in 1976-79	0.11	0.13	0.08
jobs ever changed	10.55	9.24	0.00
current tenure (weeks)	276.06	212.66	0.00
primary sector	0.23	0.08	0.00
transport	0.11	0.05	0.00
trade	0.13	0.14	0.32
finance	0.04	0.06	0.00
service, entertainment, business	0.12	0.12	0.79
proffesional services	0.11	0.30	0.00
public administration	0.05	0.05	0.75
yrs in current industry	3.56	2.62	0.00
Occupation: managerial	0.23	0.26	0.02
sales	0.07	0.08	0.02
clerical	0.06	0.21	0.00
service	0.10	0.16	0.00
farmer or laborer	0.28	0.04	0.00
machine operator	0.17	0.06	0.00
yrs in current occupation	2.18	1.73	0.00
worked full-time	0.85	0.60	0.00
weeks employed total	661.87	560.41	0.00
weeks unemployed total	62.32	49.74	0.00
bad health no work	0.05	0.05	0.07
yrs absent due to bad health since 1979	0.33	0.56	0.00

Potential confounders C related to factors that pre-date an individual's birth include **religion, race, birth order, parental education, parental place of birth** and **year of birth**. Huber and Solovyeva (2020) also acknowledge that further confounders not available in this dataset but correlated with W, X, and/or Y exist. The Table 2 summarizes the mean differences along the confounders C.

Table 2: Pre-birth confounder means for both sexes, Source: Author's replication based on Huber and Solovyeva (2020)

Variables	Men	Women	p-value of the difference
Race:black (ref: hispanic)	0.29	0.30	0.41
white	0.52	0.52	0.84
not religious (ref: catholic)	0.04	0.03	0.03
protestant	0.50	0.50	0.96
other religion	0.10	0.11	0.04
born in US	0.94	0.94	0.54
mother born in US	0.88	0.90	0.10
mother education highschool	0.39	0.37	0.05
mother's education some college	0.09	0.09	0.62
mother's education college or more	0.08	0.07	0.41
father born in US	0.88	0.88	0.41
father education highschool	0.29	0.30	0.56
father's education some college	0.09	0.08	0.10
father's education college or more	0.13	0.12	0.08
order of birth	2.98	3.07	0.14
age in 1979	17.50	17.61	0.05

4 Methodology

4.1 Gender as an Exogenous "Treatment"

This setup builds on the discussion in the literature review, yet requires some clarifications. First, I specify that treatment effect refers to the difference in outcome in the two states, however one defines the two states. One can safely replace the treatment variable with the gender variable, and the logic still holds. The ATE now is the difference in outcomes between men and women, i.e. gender gap. Note that this refers to the total gap, not the discrimination/ bias. To understand how it is deconstructed into indirect

and direct effect, I present the DAG in Figure 3.

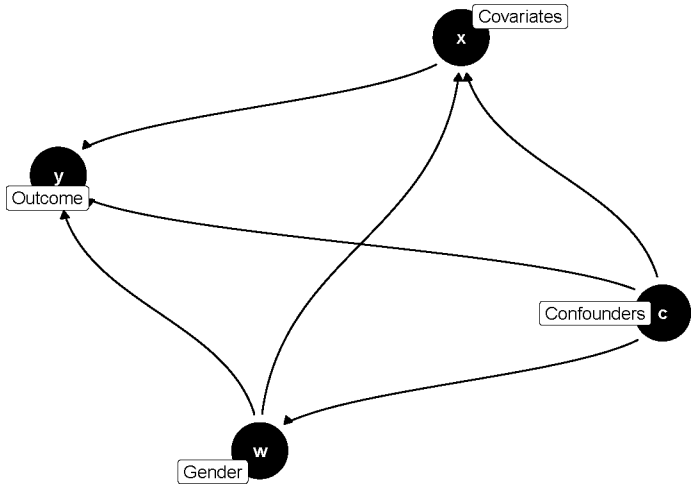

Figure 3: A causal graph of the set up. Source: Author's creation based on Huber and Solovyeva (2020)

Where Y is the variable of interest - wage, W is the gender binary variable, X is a set of mediators (covariates) that pass on the influence of gender on the wage, and C is a set of confounders pre-dating the gender binary. The logic is as following: Gender affects the wages not only directly (perception of the employer), but also through decisions made over the lifetime conditional on the gender. The latter **indirect effect** is more convoluted but actually recoverable, while the **direct effect**, can be interpreted as the discrimination.

The definition of the "treatment variable" here is neither the subject of the modern-day discourse on gender transformation, and nor should its infeasibility be a concern in this analysis, because it is the perception of the employer about the applicant's gender and gender-job compatibility that defines the causal effect. Ultimately, I am interested in comparing similar men and women, and along which paths their wages differ the most and how. One can argue about the nature of the counterfactual world, and the ideal scenario being the one with "non-discriminatory wage" which is neither male nor female wage. The philosophical discussion of the reference wages is also beyond the scope of this thesis, and better addressed by Słoczyński (2013), among others.

Altogether, this is the kind of causality without manipulation. Since the potential

outcomes framework (or at least its proponents) implied that W is the manipulated treatment assignment, this sort of a question would not be deemed causal with gender as W (Pearl 2009). This would result in leaving out numerous interesting questions unanswered. Yet few focused in exactly defining this kind of causality, ignoring the thesis that there is no causality without manipulation. Because, as Pearl (2009) puts it, the nature is in itself a great mechanism that senses values of some variables, and determines what value others take, with no need for human interference.

Notwithstanding, there are background characteristics that affect a person's wages that predate their birth. That includes their race, parental education, socio-economic status, societal norms, and etc. Together those variables likely have enough power to explain why individuals choose one career path over the other, or obtain higher education conditional on their gender. This is the set of counfounders C, that allows us to secure unconfoundedness. Ideally, the C should incorporate all the possible background variables that can affect one's further life-decisions. However because I am following Huber and Solovyeva (2020), the set is limited to what the original authors had access to.

The DAG above can also be summarized with a convenient partially-linear model:

$$
\begin{aligned}
Y_i &= \tau_i W_i + m(X_i) + g(C_i) + \epsilon_i \\
W_i &= e(W = 1|C_i) + \xi_i \\
X_i &= f(C_i) + \beta W_i + \eta_i
\end{aligned}
$$

Where the functions $m(\cdot), g(\cdot), f(\cdot), e(\cdot)$ are fully non-parametric. This representation allows me to define the impact of the large set of variables more flexibly, and view them as nuisance components. The assumptions needed to decompose the gap into causal direct and indirect effect, as stated in Huber and Solovyeva (2020) are the following:

$$
\begin{aligned}
\{Y(W, X),\ X(W)\} &\perp W \mid C \\
Y(W, X) &\perp X(W) \mid W,\ C \\
0 < e(W = 1|C) &< 1
\end{aligned}
$$

The first assumption above is **unconfoundedness**, also referred to as **"sequential conditional independence"** in the literature of mediated effects (Pearl 2012; Huber and Solovyeva 2020). That is, given a rich enough set of important pre-birth

characteristics C, a person's wage, and life decisions are independent of the gender. This is to say that the employer's perception of gender is formulated by the societal norms, among other confounders, and for men or women with the same confounders any deviation in outcomes should be random. The second assumption is similar, and together with unconfoundedness implies that after partialling out confounders, gender does not have explanatory power over the life decisions of individuals, and hence the potential wages are independent of these conditional choices. The third is the **overlap** assumption, requiring the individuals to have a comparable unit based on the propensity score $P(W = 1|C)$, where $W = 1$ refers to a male gender and $W = 0$ to female.

These assumptions allow one to partialling-out the confounders C_i as following.

$$\tilde{X}_i = X_i - \hat{f}^{-i}(C_i)$$
$$\tilde{Y}_i = Y_i - \hat{g}^{-i}(C_i)$$
$$\tilde{Y}_i = \tau_i W_i + m(\tilde{X}_i) + \tilde{\epsilon}_i$$

Overall, this is how the assumptions of unconfoundedness, ignorability, overlap and SUTVA translate into the setting with mediated effects. To further understand the intricacies of it, I define total effect as a composition of direct and indirect (mediated) effects.

4.2 Direct and Indirect Effects

Huber and Solovyeva (2020) use several methods in their work to compare how the gender gap estimates change with their specifications. The authors start with Oaxaca-Blinder decomposition, and then relax the linearity assumption using inverse propensity weights (IPW). The expressions to estimate the direct and indirect effects of gender are derived from the **"mediation formula"** of Pearl (2012):

$$ATE = E[Y_1 - Y_0] = E[\ Y(W = 1, X(W = 1)) - Y(W = 1, X(W = 0))\ ] +$$
$$+E[\ Y(W = 1, X(W = 0)) - Y(W = 0, X(W = 0))\ ]$$

Essentially, the first expectation on the right-hand side is the indirect effect of comparing the wages of men with characteristics X that are more common among men,

against men with Xs more similar to those of women. The second term, in turn, is the expression for comparing wages of men with Xs similar to women's to those of women with those characteristics. While the first difference is about how gender plays through one's life decisions with a ripple effect, the second hints at discrimination. Hence, it also corresponds to the unexplained part of the Oaxaca-Blinder decomposition (Huber and Solovyeva 2020).

For non-parametric identification Huber and Solovyeva (2020) use IPW with following formulas for each effect:

$$IE = E\left[\frac{Y\ W}{Pr(W=1|C)}\right] - E\left[\frac{Y\ W}{Pr(W=1|X,\ C)}\ \frac{1-Pr(W=1|X,C)}{1-Pr(W=1|X)}\right]$$
$$DE = E\left[\frac{Y\ W}{Pr(W=1|X,C)}\ \frac{1-Pr(W=1|X,\ C)}{1-Pr(W=1|X)}\right] - E\left[\frac{Y(1-W)}{1-Pr(W=1|C)}\right]$$

However, their analysis primarily focuses on estimating the magnitude of the gap, which may create the impression that the gap remains uniform across the sample or population with different characteristics. To gain a more comprehensive understanding of variations in the gap, it is crucial to examine the heterogeneity of the estimates. This becomes particularly significant in optimal policy learning, as it enables us to identify and target more susceptible groups, leading to improved intervention outcomes when compared to indiscriminate interventions.

To address this, I employ Causal Forests, a fully non-parametric technique designed to identify heterogeneity within subsamples (Wager and Athey 2018; Athey, Tibshirani, and Wager 2019). By utilizing doubly-robust estimation, we can effectively leverage areas where the gap is larger. These findings enable identification of specific target groups for further policy interventions.

Clearly, the expression for direct effect as put forward by Pearl (2012) can be recognized as the main output of the CF algorithm. This is not unexpected, because by design the algorithm identifies the subgroups in which conditional of covariates, the treatment effect is constant. These subgroups also have to be formed fulfilling the common support requirement, hence resulting in the expression allows me to see the difference in wages between men and women from the same backgrounds, that is not mediated through the life decisions.

5 Results

By the nature of the CF algorithm, the estimates of between-group differences at each leaf are the **direct effects** specified above. This is because a tree would split on the covariates such that the treatment effect is constant in each leaf. That is, in each leaf, conditional on X, the difference between groups is unaffected by it, considering region of common support along the propensity scores.

5.1 Exposition: A Causal Tree

First, I try to grow a causal tree (described earlier). This serves us well for interpretation and facilitates the grasp of the forest as a tree-based method. Trees are easier to interpret, because we can visualize the splits and see the estimates in the final nodes clearly. Whereas a forest is an ensemble of thousands of those trees, each (potentially) independent of another, which makes visualizing a forest extremely complicated. For further purposes, a good rule of thumb to visualize a forest is as following: Consider a new observation being taken by a tree. Based on the values of the covariates it is classified into a leaf, and the estimate for that unit is the average of the leaf. This occurs across large number of trees, each having different splitting rules and hence different leafs for the same observation. Finally, the estimates of all those trees are averaged over corresponding weights.

The tree displayed in Figure 4 originally contained 64 splits, which were pruned through cross-validation to reduce complexity and avoid overfitting. This is for example purposes, and does not incorporate propensity weighting based on confounders C. Yet still one can see the clear interpretation of the leaves, suggesting farmer or laborer women in agriculture/ manufacturing face a gap of 43%, while those not in agriculture/ manufacturing experience only 20%. Taking this further, I proceed with a proper analysis with Causal Forests.

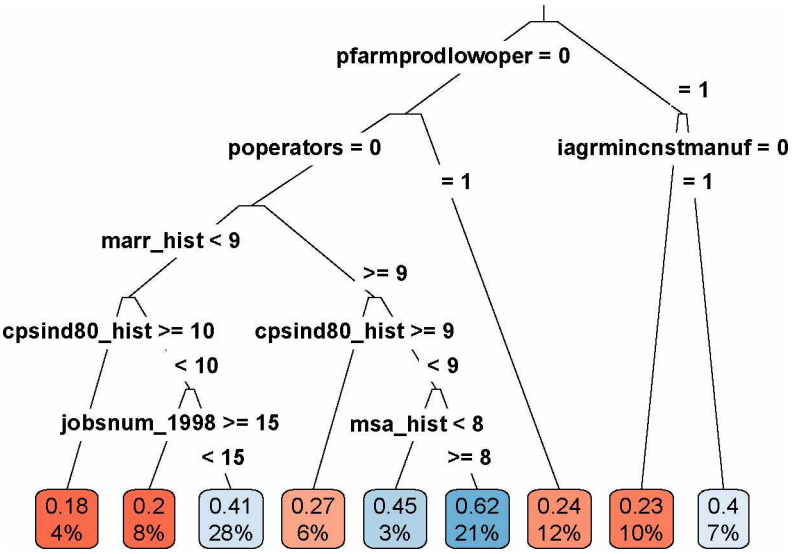

Figure 4: An example of a causal tree. This model is for exposition purposes only, as it does not control for confounders C. It is pruned, however, as the original tree had 64 splits. This is to show how one can cope with overfitting with trees. Source: Author's calculations

5.2 Direct Effect or the Gender Wage Gap

Running several versions of CF, I start with the specification where gender is taken as random, not controlling for any confounders, and then add the controls and confounders. The algorithm can take the propensity scores as an external input, but if not explicitly given, the nuisance components are calculated automatically. In a randomized assignment settings, one can input the share of treated as a propensity score. However, in my setup, I ensure the propensities are calculated from the confounders, and then input them to the final Causal Forest. The "proper" specification is the one that takes estimates $\hat{e}(C)$ and also partials out the C from the Y and X (i.e $\mu(X, W) = E[Y|X, W, C]$). I present the results below.

Augmented Inverse-Propensity Weighted (AIPW) ATE is the recommended way to compute average treatment effects in observational data (Tibshirani et al. 2022). It consists of averaging the doubly robust scores, where $\tau^{-i}(X_i)$ and $e^{-i}(C_i)$ are out-of-bag estimates. The I_q stands for all the individuals within the given quantile of the effects (I display this later in figure 6).

$$\frac{1}{|I_q|}\sum_{i\in I_q}\hat{\tau}^{-i}(X_i)+\frac{W_i-\hat{e}^{-i}(C_i)}{\hat{e}^{-i}(C_i)\left(1-\hat{e}^{-i}(C_i)\right)}\left(Y_i-\hat{\mu}_{W_i}(W_i,X_i)\right)$$

Table 3: Comparison of ATE estimates from different implementations of CF, Source: Author's calculations

Estimates	Standard error	Specification
0.205	0.015	Baseline CF only with covariates X, gender as random
0.218	0.018	CF with automatic propensity estimation, X only
0.196	0.015	CF with propensity scores over confounders W
0.215	0.018	Kitchen-sink CF (both X and W)
0.208	0.015	Proper CF

Compared to IPW with W estimates of Huber and Solovyeva (2020), I recover a larger direct gender wage gap. My point-estimate is 0.208 with smaller standard errors 0.015, which is within the confidence interval of the estimates of Huber and Solovyeva (2020), they recover 0.171 with standard error 0.042. This is an expected difference, since the methods we use are yet both non-parametric, still quite dissimilar. Moreover, the standard errors I report are based on out-of-bag estimates of the forest.

Because I have doubly-robust estimates of the gender gap and also correct for the sample biases by using out-of-bag nuisance functions in validation, we can more confidently predict individual treatment effects based on the forest. In Figure 5 I display the distribution of the individual treatment effects. The graph shows the direct gap estimated for each set of individual characteristics X, *and* C present in the sample.

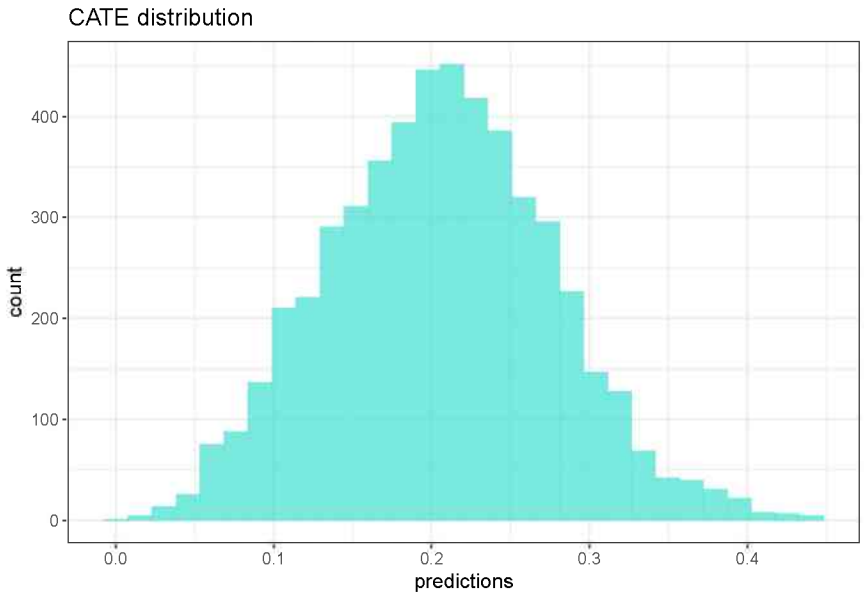

Figure 5: Distribution of estimated individual treatment effects, or the individual gender gap. Source: Author's calculations.

5.3 Quantiles of the Gender Wage Gap and RATE

I also check the difference between quantiles of treatment effects in Figure 6, as mentioned above, to gain a clearer understanding of the heterogeneity in the effects than on a basic histogram. Note that the average estimates of the treatment effect that is obtained by averaging doubly-robust scores may not be monotonic. That is, the average estimate for group Q3 may end up being smaller than the one for Q2. Asymptotically, these differences should disappear, but this is a common occurrence in small samples.

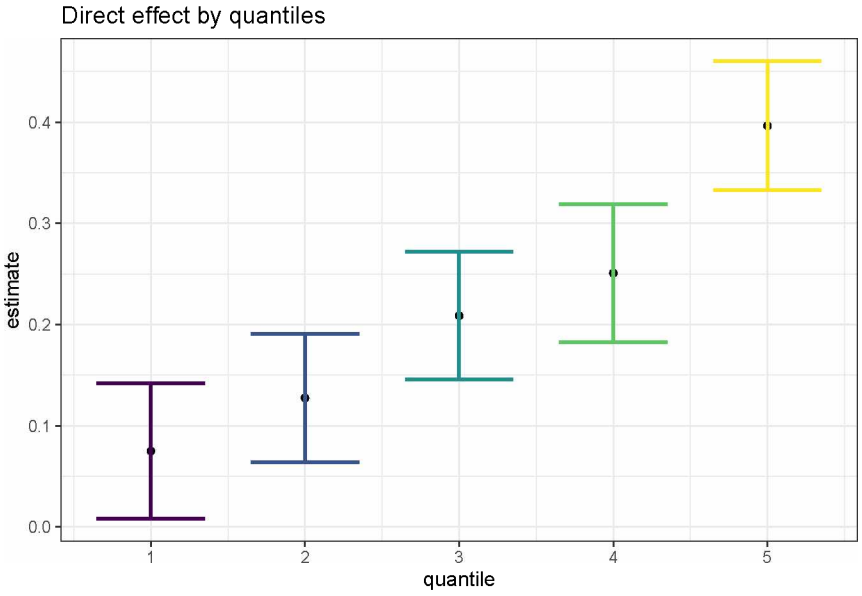

Figure 6: The estimates for gender gap accord to the quantiles. Source: Author's calculations.

Another way to look at this is the more recently developed **Ranked ATE (RATE)** by Yadlowsky et al. (2021). RATE essentially allows one to see if there is heterogeneity in the estimates by comparing each of the percentiles of the treatment effect to the ATE, and then ranking those differences. This is very similar to the quantiles approach described above, but more uniform and informative. Also, by calculating the area under the RATE curve, one would get the total gains from implementing an intervention, for p-th percentile of more susceptible population versus extending it to everyone. If there are HTEs, the curve will start high for the individuals with the highest expected benefit and decline until it equals ATE when q=1, i.e., everyone is included.

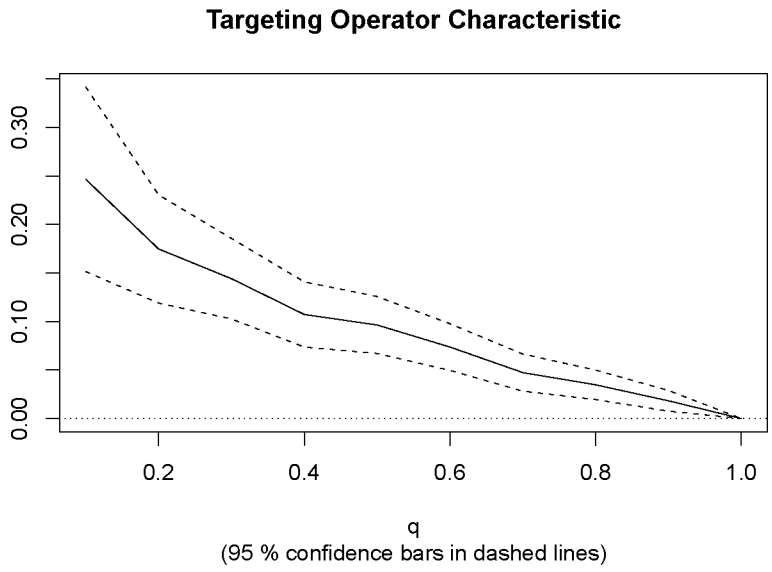

Figure 7: Ranked Average Treatment Effects (RATE). Source: Author's calculations.

From the Figure 7 it is clear that the gender gap is clearly uniformly positive (in favor of men) and there is heterogeneity in how it is manifested. Again, the vertical axis is the difference between that percentile gender gap and the average gender gap.

5.4 Propensity Scores

As Huber and Solovyeva (2020) assume unconfoundedness, and because it plays a pivotal role in identification, I make sure to check if it holds. As outlined previously, controlling for confounders, the outcome and covariates are independent of gender, thus, I estimated propensity scores to reweigh based on them. This should give more comparable individuals based on their pre-birth characteristics. Subsequently, I ensure that I have enough comparable units across my estimated propensity scores by checking the overlap region.

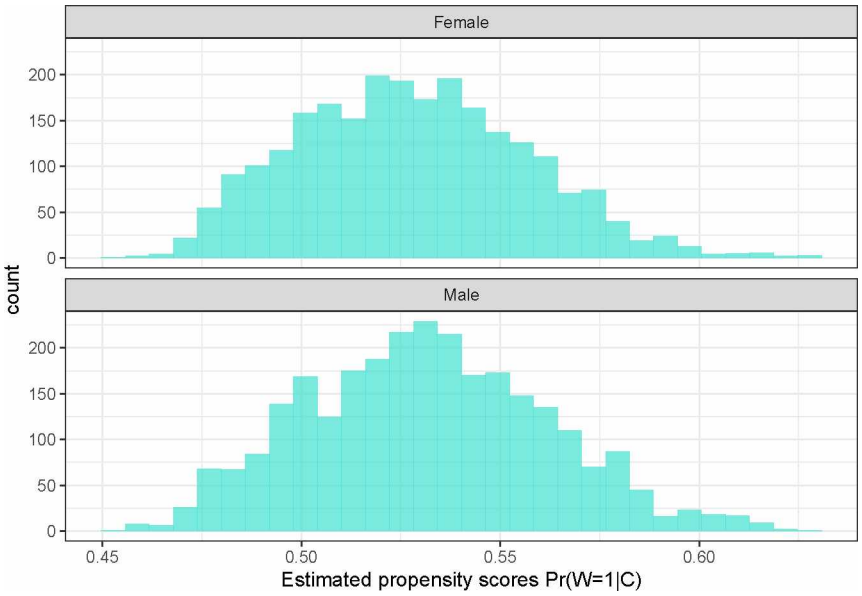

Figure 8: Propensity scores Pr(W=1|C). These are estimated with a separate regression forest and then fed to the final CF. Source: Author's calculations.

The scores on the Figure 8 being bounded away from extremes indicate that male and female composition is comparable, and there is a reasonable overlap region. This is in line with what Huber and Solovyeva (2020) obtained with their propensities. All the estimates reported in this section are based on that region of common support.

5.5 An Omnibus Test for Heterogeneity

As the CF uses Robinson (1988)' orthogonalized loss function as an objective, it can also serve as a back-test to check how well the forest has performed, as following:

$$Y_i - \hat{\mu}^{-i}(W_i, X_i) = \alpha\bar{\tau}\left(W_i - \hat{e}^{-i}(C_i)\right) + \beta\left(\hat{\tau}^{-i}(X_i) - \bar{\tau}\right)\left(W_i - \hat{e}^{-i}(C_i)\right) + \epsilon$$
$$\bar{\tau} := \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}^{-i}(X_i)$$

This is essentially decomposition of the residuals into the components corresponding to treatment effect and its variance. The coefficients α and β allow me to evaluate the performance of the estimates. If $\alpha = 1$, then the average prediction produced by the forest is correct. Meanwhile, if $\beta = 1$, then the forest predictions adequately capture the underlying heterogeneity.

In addition, β is a measure of how the CATE predictions covary with true CATE.

Therefore, the p-value on the estimate of coefficient also acts as an omnibus test for the presence of heterogeneity. If the coefficient is significantly greater than zero, then we can reject the null of no heterogeneity. However, coefficients smaller than 0 are not meaningful and should not be interpreted.

Table 4: Best linear fit using forest predictions (on held-out data), with heteroskedasticity-robust (HC3) SEs.

	Estimates	Standard error	t-statistic	p-value
alpha	1.01	0.07	13.65	0
beta	1.48	0.23	6.32	0

The results suggest that the algorithm correctly identifies the constant part of the gender gap, while also uncovering significant heterogeneity along the covariates.

5.6 Heterogeneity in the Gender Gap

Thus far, I have only investigated the average estimates, and checked if there is heterogeneity present in the gender gap. Now I turn to show where exactly the variation in the gap originates from. It is important to keep in mind that all the estimates of the forest are the direct effects, so they can be interpreted as gender bias under sequential identification.

5.6.1 Variable importance

First, I consider what variables the forest deemed important, i.e. along which of them trees split most often. This is a rough measure to look at, because just as with random forests, this measure is ambiguous (usually one uses the **variance inflation factor (VIF)** to reweigh the variable importance). Nevertheless, the Figure 9 is a good starting point.

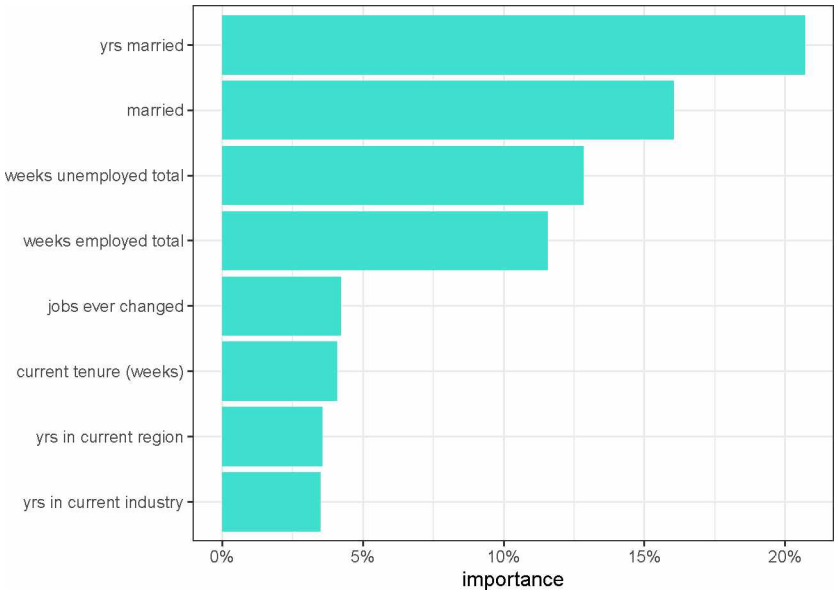

Figure 9: Variable importance of the Causal Forest, shows the variables on which splits were made most often. The actual number of variables is larger, only the most important ones displayed. Source: Author's calculations

5.6.2 Heterogeneity along the quantiles

The approach of comparing all covariates across quantiles of treatment effects presents a fuller picture of how high-treatment-effect individuals differ from low-treatment-effect individuals. In the Table 5 in the appendix I display the average values of the covariates according to each quantile of the direct effect. The actual table is quite large, so I include visualizations of only those variables that have significant changes in the means for each treatment effect quantile in the Figure 10.

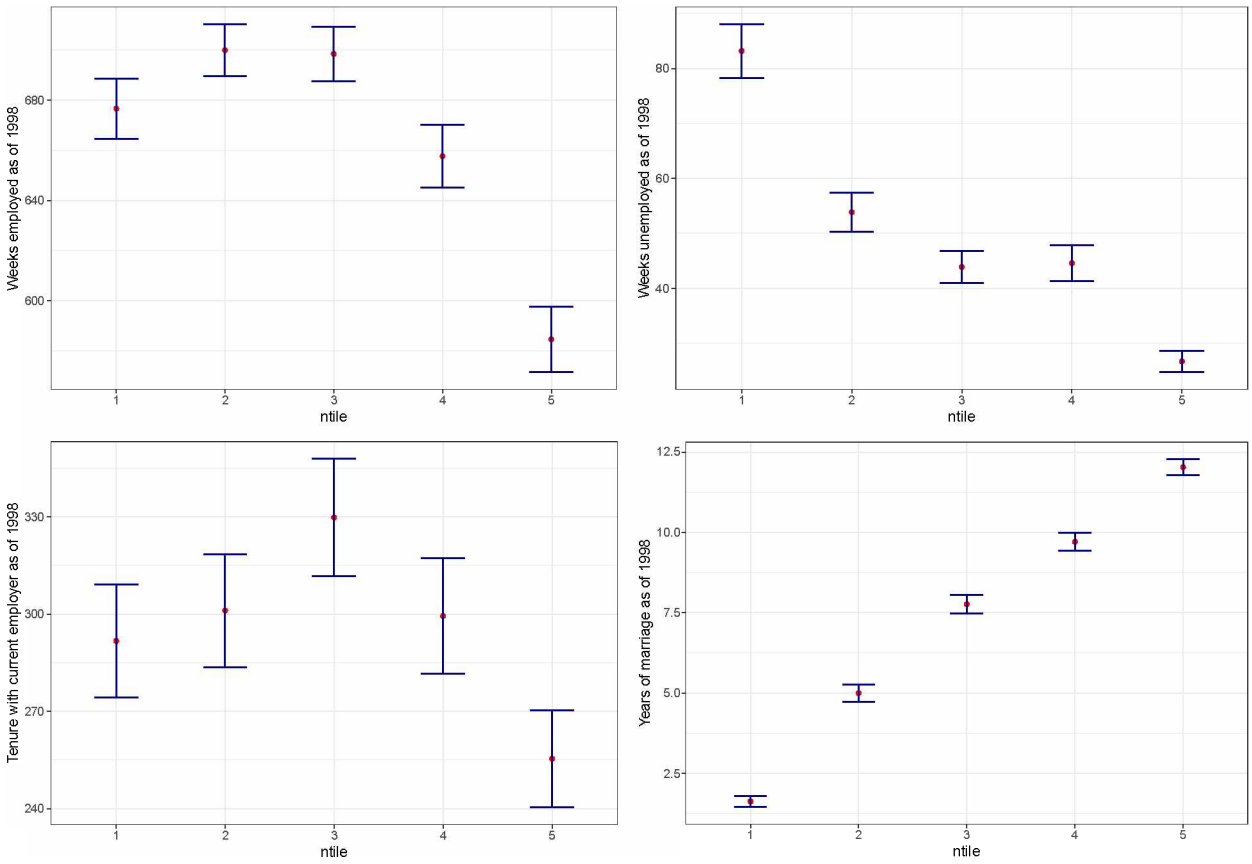

Figure 10: Partial dependence plots to see how average characteristics differ among the quantiles presented above. Source: Author's calculations.

5.6.3 Partial dependence

In traditional heterogeneity estimation using generalized linear models, understanding the interrelations of variables is relatively straightforward as one can easily examine them visually. However, with ensemble machine learning models, interpreting the effect of variables on the outcome becomes more challenging due to their black-box nature. To address this, the SHAP (SHapley Additive exPlanations) method provides a powerful tool for interpreting such models by summarizing the meaningful value-added by each variable. This approach allows us to report the average contribution of each variable to the outcome, similar to parameter values in a linear regression (Lundberg and Lee 2017). While SHAP is effective, it can be computationally intensive for certain models, and there is currently no available wrapper for the Causal Forest of the grf package.

To overcome this limitation, I employ an alternative method that is computationally

lightweight and easy to implement. I present a set of more meaningful plots by fixing all other covariates at their medians and allowing variation only along one of them. This approach enables the traditional ceteris paribus interpretation, which helps us understand the impact of individual variables while holding other factors constant. However, it is essential to note that the error bars in these plots might be wide since occurrences of these median values for other covariates could be rare in the data. Despite this limitation, this approach provides valuable insights into the heterogeneity of treatment effects in the Causal Forest framework.

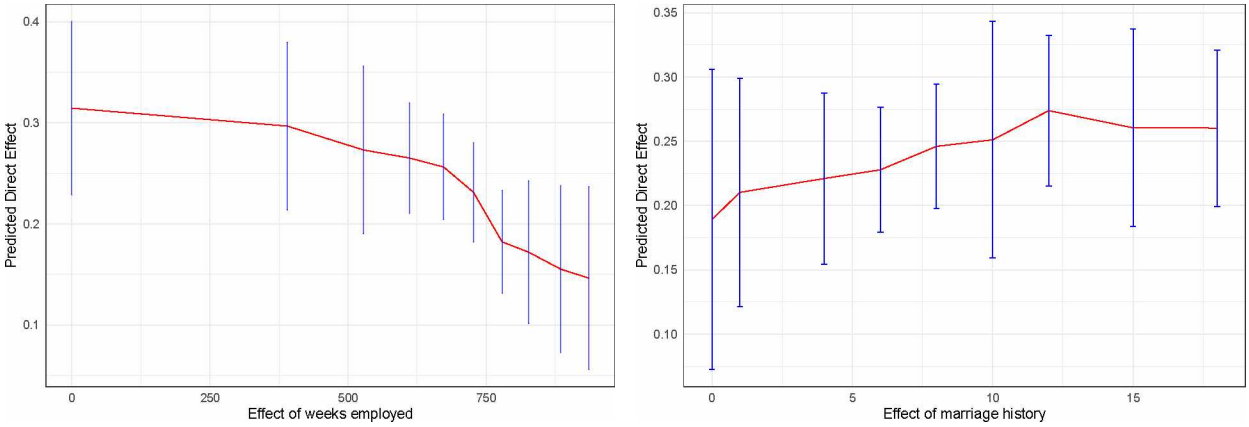

Figure 11: Evaluating the variation in gender gap by quantiles of experience (left) and marriage (right) as of 1998. All the other variables fixed at their medians, and hence this can be interpreted as ceteris paribus effect. The intervals are wide because not many observations were concentrated at or around the median values in the sample. Source: Author's calculations

The Figure 11 above shows that among more experienced individuals, controlling for everything else, the gap can narrow. Yet the wider confidence intervals do not provide any clarity over this observation. The scarcity of the comparable men and women along median values of covariates is even more expressive in the right plot in Figure 11. Here, although one can see the gender wage gap increasing in the years of marriage, the error bars suggest the difference can be constant.

The Figure 13 in the appendix provides a hint that the gap is monotonically decreasing as the time spent out of the labor market increases. One can observe that men and women recover differently from a job displacement. I take this observation further by allowing variation along the years in marriage too, holding everything else

constant.

This relationship presented in the Figure 12. The color scheme allows to locate the subgroups with the most severe gender wage gap experienced. Supporting the previous observation, I recover a significant difference in wages among individuals who had less than 6 weeks of unemployment in total and were married for at least 12 years. This approach to interpreting the findings of Causal Forest can directly translate into a policy rule, that for instance, targets individuals who are in long-term marriage and also have not been displaced for long.

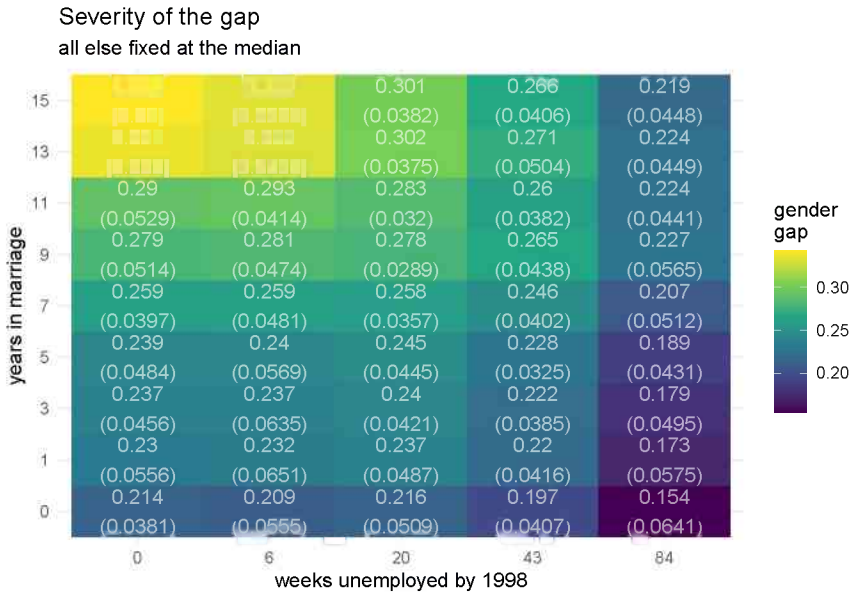

Figure 12: Variation in the gender gap along two variables: years in current region and years in marriage. All the other variables fixed at their medians. This is a more robust check of the dependence plots above. The White standard errors are reported in parentheses at each tile. Source: Author's calculations

Another such policy rule can be seen in the Figure 15 of the appendix, where I allow variation only along the marriage years and the history of residing in current region. It is important to emphasize, that in holding everything else equal, the median person in the sample lived in the South region and had less than high-school education, among other characteristics. Once again, I recover over 33% difference in wages between median men and women, who were in long term marriage and were new to the region. One can think of a support program designed to ease the adjustment of the newly-moved

women with a marriage long history.

Furthermore, for binary indicators with more concentration over the support, significant differences can be observed in the Figure 14 in appendix. It is clear that, controlling for everything else, married women experience 6 *to* 10% larger wage gap than singles. A magnitude of similar inequality 7 *to* 12% can be observed between women working in the services sector too.

6 Discussion

The results can be summarized in the following ways. First, I find the wages of men almost always higher than those of comparable women, coming from identical backgrounds. Second, I identify near-significant evidence that gender gap actually narrows with the experience (See Figure 11). Third, the gap is wider among public sector workers, and also among married women, controlling for everything else. Lastly, the gap is larger for individuals who had been unemployed less than 6 weeks before, and have a marriage history of more than 10 years. This may suggest that one can assign an audit of jobs and employee contracts among public sector workers, rather than a roll-out audit for everyone.

For long-term married women who are new to the region, the wage gap is also more pronounced, spanning over 33%. These findings can encourage targeting rules for eased mobility and integration of female workers into the particular region. These detailed results are relevant for individuals in the South region with less than highschool education, descriptive of a median person in the sample. However, it is possible to evaluate people in higher quantiles of the sample, and organize a policy relevant for them. This is possible because the counterfactuals I constructed using CF are doubly-robust in interpretation.

Although in my analysis I followed a more credible identification of the gender wage gap, and augmented the estimation of heterogenous effects with a doubly-robust estimator, there are limitations. The main limitation is the fact that I ignore the selection into labor market participation and focus on the subsample of workers who worked full-time at the time of the survey. The reason for that is that original authors find no significant evidence when using IV estimation with a child's age as an instrument

(Huber and Solovyeva 2020). Clearly, the analysis can be extended with an appropriate control strategy for selection into labor force.

Moreover, the data are cross-sectional, which only allows me to see a snapshot of people's experiences of wage gap. A better way to look at it is to exploit the time dimension, also capturing the evolution of the gap.

Third, the informativeness of the variables is somewhat limited for a specific policy rule. Hence my findings can be deemed less actionable, offering a very generic, descriptive image of the heterogeneity in gender wage discrimination. One could certainly recover heterogeneity in the gap using a richer set of confounders (e.g. including genetic factors). While the external validity of this analysis can be lacking, I nevertheless showcase the value-added of the Causal Forests in providing the tailored policy rules by exploring heterogeneity in gender wage gap.

7 Conclusion

In this thesis, my primary focus has been on estimating heterogeneity in gender bias using Causal Forests. I conducted an extensive literature review, addressing critical issues related to estimating treatment effects, such as selection bias, heterogeneity, and mediated effects. Additionally, I explored popular machine learning methods commonly employed by economists and their applicability in causal inference.

Building upon the study of Huber and Solovyeva (2020), I revisited their findings using a completely non-parametric and doubly-robust Causal Forest algorithm. I reported the results and also performed sensitivity checks, assessing the model's calibration to validate the heterogeneity estimates.

The partial dependence analysis indicated significant heterogeneity based on factors like marriage, regional history, unemployment, and public sector occupations. This suggests that older married men tend to earn significantly more than equally comparable women, while the difference diminishes for individuals with more extended periods of absence from the job market. This finding also sheds light on how married men and women recover differently after experiencing job loss. Also, I uncover that older married women can be at a significant disadvantage when newly moved to the region.

Although my results may not lead to actionable policy recommendations, the

heterogeneity analysis unmistakably reveals that, on average, men in the sample earned more than equally comparable women. The thesis aimed to demonstrate the value-added by employing causal machine learning, particularly Causal Forests, in addressing complex socio-economic phenomena. By estimating heterogeneity in the gender-wage gap and illustrating how the results can be visualized, this thesis serves as an example of utilizing gender bias variation for policy targeting. Through this work, I sought to contribute to the evolving literature of causal machine learning applications and also bridge the gap between traditional decomposition methods and the credible analysis of causal effects, thereby providing a clearer understanding of this critical issue.

References

Angrist, Joshua, and William N Evans. 1996. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size."

Angrist, Joshua, and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Arellano, Manuel, and Stéphane Bonhomme. 2017. "Quantile Selection Models with an Application to Understanding Changes in Wage Inequality." *Econometrica* 85 (1): 1–28.

Athey, Susan. 2017. "Beyond Prediction: Using Big Data for Policy Problems." *Science* 355 (6324): 483–85.

———. 2018. "The Impact of Machine Learning on Economics." In *The Economics of Artificial Intelligence: An Agenda*, 507–47. University of Chicago Press. http://www.nber.org/chapters/c14009.

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–60. https://doi.org/10.1073/pnas.1510489113.

Athey, Susan, and Guido W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2). https://doi.org/10.1214/18-aos1709.

Athey, Susan, and Stefan Wager. 2021. "Policy Learning with Observational Data." *Econometrica* 89 (1): 133–61.

Bach, Philipp, Victor Chernozhukov, and Martin Spindler. 2018. "Closing the US Gender Wage Gap Requires Understanding Its Heterogeneity." *arXiv Preprint arXiv:1812.04345*.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014a. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28 (2): 29–50.

———. 2014b. "Inference on Treatment Effects After Selection Among High-Dimensional Controls." *The Review of Economic Studies* 81 (2): 608–50.

Blake, Thomas, Chris Nosko, and Steven Tadelis. 2015. "Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment." *Econometrica* 83 (1): 155–74.

Blau, Francine D, and Lawrence M Kahn. 2017. "The Gender Wage Gap: Extent, Trends, and Explanations." *Journal of Economic Literature* 55 (3): 789–865.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and Regression Trees*. CRC press.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." Oxford University Press Oxford, UK.

Chernozhukov, Victor, Iván Fernández-Val, and Siyi Luo. 2023. *Distribution Regression with Sample Selection and UK Wage Decomposition*. Cemmap, Centre for Microdata Methods; Practice, The Institute for Fiscal Studies, Department of Economics, UCL.

Cunningham, Scott. 2022a. *Interview with Steve Tadelis, UC Berkeley Haas Business School Professor and Formerly eBay*. The Mixtape with Scott. Substack. https://causalinf.substack.com/p/interview-with-steve-tadelis-uc-berkeley-cac#details.

———. 2022b. *Interview with Susan Athey, Professor at Stanford, President of AEA*. The Mixtape with Scott. Substack. https://causalinf.substack.com/p/interview-with-susan-athey-professor-48e#details.

DiNardo, John, Nicole M Fortin, and Thomas Lemieux. 1995. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." Working Paper 5093. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w5093.

Donohue, John J, and Steven D Levitt. 2001. "The Impact of Legalized Abortion on Crime." *The Quarterly Journal of Economics* 116 (2): 379–420.

Farrell, Max H, Tengyuan Liang, and Sanjog Misra. 2021. "Deep Neural Networks for Estimation and Inference." *Econometrica* 89 (1): 181–213.

Friedberg, Rina, Julie Tibshirani, Susan Athey, and Stefan Wager. 2020. "Local Linear Forests." *Journal of Computational and Graphical Statistics* 30 (2): 503–17.

Glaeser, Edward L, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. "Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy."

American Economic Review 106 (5): 114–18.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica*, 315–31.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. Springer.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica: Journal of the Econometric Society*, 153–61.

Heckman, James J, Hidehiko Ichimura, Jeffrey A Smith, and Petra E Todd. 1998. "Characterizing Selection Bias Using Experimental Data." National bureau of economic research Cambridge, Mass., USA.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Huber, Martin. 2019. "Replication data for 'On the sensitivity of wage gap decompositions'." Harvard Dataverse. https://doi.org/10.7910/DVN/B5HWTZ.

Huber, Martin, and Anna Solovyeva. 2020. "On the Sensitivity of Wage Gap Decompositions." *Journal of Labor Research* 41: 1–33.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116 (10): 4156–65.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review*, 604–20.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *The American Economic Review* 73 (1): 31–43.

Lechner, Michael. 2023. "Causal Machine Learning and Its Use for Public Policy." *Swiss Journal of Economics and Statistics* 159 (1): 1–15.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. https:

//doi.org/10.1257/jep.31.2.87.

Neyman, Jerzy. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles." *Ann. Agricultural Sciences*, 1–51.

Nie, Xinkun, and Stefan Wager. 2021. "Quasi-Oracle Estimation of Heterogeneous Treatment Effects." *Biometrika* 108 (2): 299–319.

Pearl, Judea et al. 2000. "Models, Reasoning and Inference." *Cambridge, UK: CambridgeUniversityPress* 19 (2): 3.

Pearl, Judea. 2009. *Causality*. Cambridge university press.

———. 2012. "The Causal Mediation Formula—a Guide to the Assessment of Pathways and Mechanisms." *Prevention Science* 13: 426–36.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66.

Robinson, Peter M. 1988. "Root-n-Consistent Semiparametric Regression." *Econometrica: Journal of the Econometric Society*, 931–54.

Roy, Andrew Donald. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3 (2): 135–46.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688.

Słoczyński, Tymon. 2013. "Population Average Gender Effects."

Smith, Jeffrey A, and Petra E Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1-2): 305–53.

Tibshirani, Julie, Susan Athey, Erik Sverdrup, and Stefan Wager. 2022. *Grf: Generalized Random Forests*. https://CRAN.R-project.org/package=grf.

Van Der Laan, Mark J, and Daniel Rubin. 2006. "Targeted Maximum Likelihood Learning." *The International Journal of Biostatistics* 2 (1).

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42. https://doi.org/10.1080/01621459.2017.1319839.

Yadlowsky, Steve, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. 2021. "Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects." *arXiv e-Prints*, arXiv–2111.

Yamaguchi, Kazuo. 2015. "Decomposition of Gender or Racial Inequality with Endogenous Intervening Covariates: An Extension of the DiNardo-Fortin-Lemieux Method." *Sociological Methodology* 45 (1): 388–428.

Appendix

Table 5: Average covariate values in each quantile

Covariates	Bottom 20%	20%-40%	40%-60%	60%-80%	Top 20%
married	0.043	0.397	0.693	0.853	0.962
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
yrs married	1.619	4.994	7.763	9.707	12.03
	(0.132)	(0.132)	(0.132)	(0.132)	(0.132)
North East	0.17	0.163	0.154	0.151	0.112
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
North central	0.168	0.245	0.274	0.282	0.25
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
West	0.188	0.186	0.187	0.182	0.218
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
yrs in current region	15.54	15.47	15.24	14.92	14.27
	(0.116)	(0.116)	(0.116)	(0.116)	(0.116)
SMSA	0.844	0.843	0.813	0.771	0.791
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
yrs in current SMSA	14.48	14.2	13.98	13.65	12.97
	(0.132)	(0.132)	(0.132)	(0.132)	(0.132)
Degree highschool	0.464	0.466	0.43	0.428	0.363
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
some college	0.249	0.217	0.232	0.245	0.31
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
college or more	0.172	0.226	0.257	0.233	0.258
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
first job before 1975	0.047	0.057	0.05	0.059	0.061
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
first job in 1976-79	0.149	0.123	0.113	0.109	0.113
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Covariates	Bottom 20%	20%-40%	40%-60%	60%-80%	Top 20%
jobs ever changed	11.01	10.89	9.871	9.649	8.764
	(0.173)	(0.173)	(0.173)	(0.173)	(0.173)
current tenure (weeks)	291.7	301.1	329.8	299.4	255.4
	(8.756)	(8.756)	(8.756)	(8.76)	(8.76)
primary sector	0.202	0.187	0.163	0.164	0.123
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
transport	0.111	0.113	0.097	0.072	0.07
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
trade	0.14	0.122	0.144	0.156	0.149
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
finance	0.053	0.059	0.076	0.047	0.064
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
service, entertainment, business	0.136	0.122	0.12	0.122	0.121
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
proffesional services	0.214	0.216	0.208	0.233	0.274
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
public administration	0.051	0.056	0.052	0.073	0.083
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
yrs in current industry	4.386	4.019	4.276	3.389	2.401
	(0.125)	(0.125)	(0.125)	(0.125)	(0.125)
Occupation: managerial	0.236	0.26	0.299	0.306	0.33
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
sales	0.06	0.086	0.08	0.083	0.086
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
clerical	0.151	0.16	0.126	0.145	0.167
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
service	0.148	0.103	0.109	0.137	0.162
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
farmer or laborer	0.231	0.172	0.182	0.156	0.125
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)

Covariates	Bottom 20%	20%-40%	40%-60%	60%-80%	Top 20%
machine operator	0.133	0.169	0.135	0.128	0.073
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
yrs in current occupation	2.814	2.511	2.373	2.092	1.79
	(0.095)	(0.095)	(0.095)	(0.095)	(0.095)
worked full-time	0.872	0.886	0.89	0.85	0.819
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
weeks employed total	676.6	699.8	698.4	657.6	584.6
	(5.995)	(5.995)	(5.995)	(5.998)	(5.998)
weeks unemployed total	83.19	53.86	43.91	44.6	26.72
	(1.754)	(1.754)	(1.754)	(1.755)	(1.755)
bad health no work	0.007	0.007	0.005	0.003	0.009
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
yrs absent due to bad health since 1979	0.14	0.146	0.173	0.2	0.264
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)

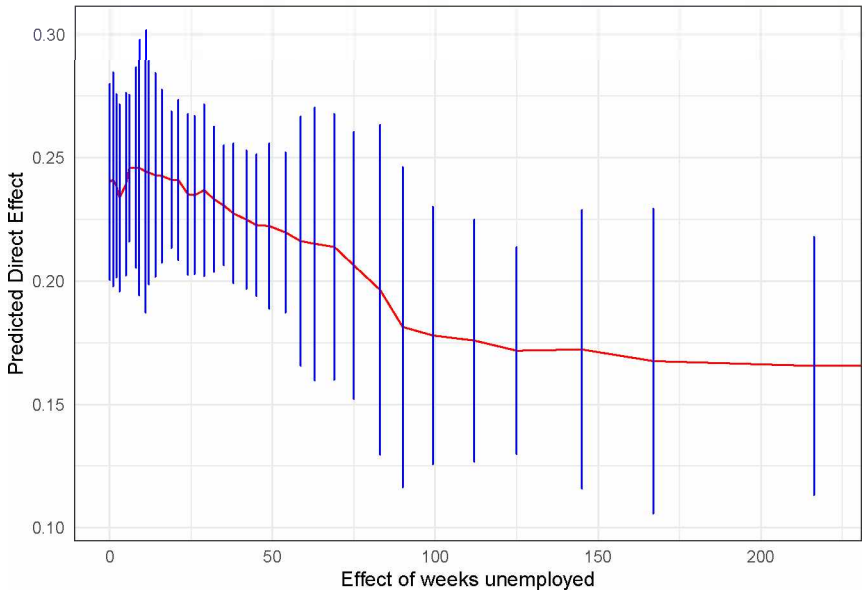

Figure 13: Evaluating the variation in gender gap by quantiles of weeks spent unemployed as of 1998. All the other variables fixed at their medians, and hence this can be interpreted as ceteris paribus effect. The intervals are wide because not many observations were concentrated at or around the median values in the sample. Source: Author's calculations

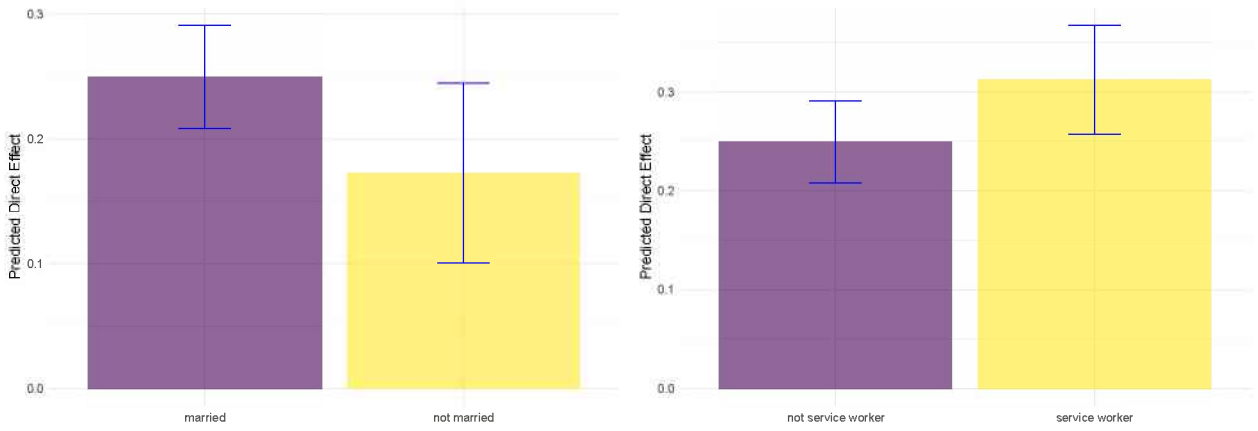

Figure 14: Evaluating the variation in gender gap by for binary variables of public service employment and marital status. All the other variables fixed at their medians, and hence this can be interpreted as ceteris paribus effect. The intervals are wide because not many observations were concentrated at or around the median values in the sample. Source: Author's calculations

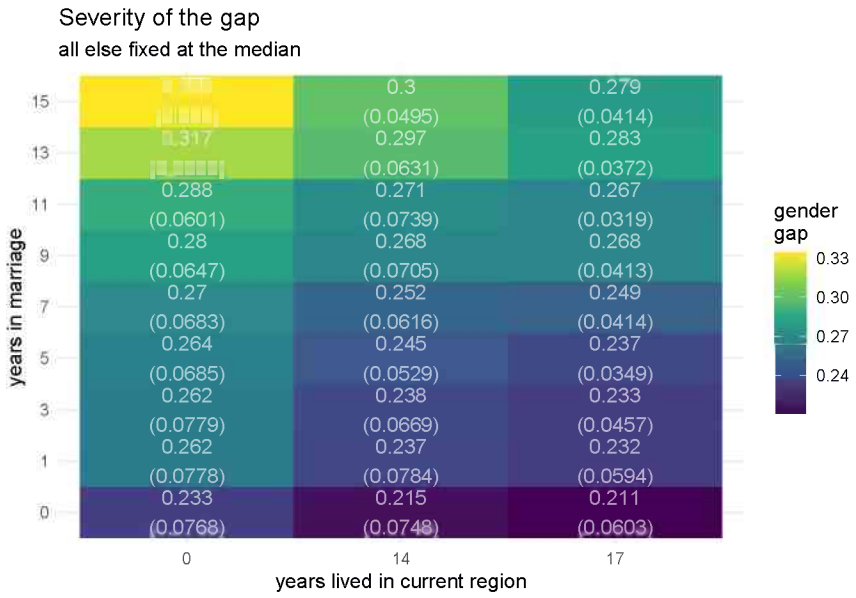

Figure 15: Variation in the gender gap along two variables: years lived in the current region and years in marriage. All the other variables fixed at their medians. This is a more robust check of the dependence plots above. The White standard errors are reported in parentheses at each tile. Source: Author's calculations