Detecting Heterogeneity in Labor Market Discrimination

An Application of Causal Forests to the Oreopoulos (2011) Dataset

Stephen Min

Simon Fraser University

Abstract

This paper applies the causal forest machine learning algorithm to detect heterogeneity in labor market discrimination in Canada using data from the resume correspondence study by Oreopoulos (2011). In contrast to the original study, my analysis reveals that the effect of having an English name on callback rates varies across applicant characteristics. Notably, the effect of an English name on callback rates was 4 percentage points higher for resumes with feminine names compared to those with masculine names. Conversely, possessing a degree from a top-ranked university and listing extracurricular skills appear to mitigate this effect, though with less certainty.

Contents

1 Introduction

Labor market discrimination against immigrants and ethnic minorities remains a persistent concern in many developed countries despite decades of equal opportunity legislation and changing social norms. This discrimination often manifests in subtle ways, such as differential treatment based on perceived ethnicity as signaled by an applicant's name. A growing body of research has sought to quantify and understand this phenomenon through various experimental methods. In the context of the Canadian labor market, Oreopoulos (2011) is perhaps the most notable study in its depth and scope. Using a large-scale resume correspondence study, Oreopoulos demonstrated significant differences in callback rates for job applications with English versus foreign-sounding names, providing compelling evidence of name-based discrimination. This research, along with similar studies in other contexts (Bertrand and Mullainathan, 2004; Carlsson and Rooth, 2007; Pager, 2003), has been instrumental in highlighting the ongoing challenges faced by ethnic minorities in the job market.

While the existing literature on discrimination has provided valuable insights, it is limited by the methods that were available at the time. These methods often require imposing strong assumptions about the functional form of relationships between variables, potentially obscuring important nuances in how discrimination manifests. For instance, many existing studies commonly assume that the outcome of interest (e.g. callback rates) is a linear function of the relevant variables[1], and the treatment effect varies through interaction terms that must be specified by the researcher. The underlying models of existing studies could be misspecified, and the impact of name-based discrimination could vary in an unknown way based on factors such as gender, education level, or other resume characteristics (Heckman, 1998).

Recent advancements in machine learning offer a promising alternative to traditional methods via nonparametric estimation. Athey and Imbens (2019) provide an overview of machine learning methods that have been proven valuable to econometricians. For the specific case of treatment effect estimation, notable work has been done to incorporate the usage of machine learning algorithms into causal inference (Chernozhukov et al., 2018; Nie and Wager, 2021).

This paper builds on Oreopoulos (2011) by employing the *causal forest* algorithm (Athey et al., 2019), a recent development in causal machine learning that allows for non-parametric estimation of heterogeneous treatment effects. My findings reveal detectable heterogeneity in the effect of having an English name on callback rates. Notably, I find strong evidence that having a feminine name amplifies the positive impact of having an English name, suggesting a compounding effect of perceived gender and perceived ethnicity in hiring decisions. Conversely, possessing a degree from a high-quality university[2] and listing extracurricular skills on a resume appear to mitigate the effect of having an English name. I validate my findings through a robustness check, correcting for multiple hypothesis testing with the Romano-Wolf procedure. While this adjustment pushes p-values out of conventional significance levels, the resulting values remain low enough to warrant serious consideration of the observed heterogeneity[3].

This research makes two main contributions to the literature on discrimination in the labor market. First, it provides a more nuanced understanding of labor market discrimination by revealing how its effects vary across different subgroups. Second, my methodological approach demonstrates the potential of causal machine learning to uncover otherwise hard-to-detect patterns in complex social phenomena, providing a blueprint for future investigations in labor economics and beyond.

The remainder of this paper is structured as follows: section 2 presents the theoretical framework I operate under, including a description of the causal forest algorithm, and the empirical strategy for analyzing heterogeneity; section 3 describes the dataset, originating from Oreopoulos (2011); section 4 presents my main results; section 5 discusses some implications of my findings and concludes.

[1] Or that a transformation can be made to make the method valid.
[2] This is defined in the original paper by Oreopoulos. I also include the definition in section 3.
[3] Because of this, the robustness check is placed in appendix A.

2 Methodology

2.1 Theoretical Framework

Given some independent and identically distributed data with observations (X_i, Y_i, W_i) for $i = 1, \ldots, n$ where $X_i \in \mathcal{X}$ are individual covariates, $Y_i \in \mathbb{R}$ is the observed outcome, and $W_i \in \{0, 1\}$ is the treatment assignment, we consider the existence of potential outcomes $Y_i(1)$ and $Y_i(0)$ corresponding to the outcome that would have been observed given the treatment assignment of $W_i = 1$ or 0 respectively. We are interested in estimating the effect of $W_i = 1$ on the outcome without needing to specify a strict functional form. That is, we posit the following model:

$$Y_i = \tau(X_i)W_i + f(X_i) + \epsilon_i, \quad \mathbb{E}\left[\epsilon_i \,\middle|\, W_i, X_i\right] = 0 \tag{1}$$

and interpret, for $x \in \mathcal{X}$, the estimate $\hat{\tau}(x)$ as the conditional average treatment effect (CATE)[4]:

$$\hat{\tau}(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right].$$

To properly identify the CATE, we must make three assumptions. The necessity of these assumptions is explained in appendix B.

Assumption 1 (Unconfoundedness)**.**

$$Y_i(1), Y_i(0) \perp W_i \,\middle|\, X_i$$

i.e. potential outcomes are independent of treatment assignment given covariates.

Assumption 2 (Overlap)**.**

$$0 < \Pr[W_i = 1 \,|\, X_i = x] < 1 \quad \forall x \in \mathcal{X}$$

i.e. there is a positive probability of receiving treatment for all covariate values.

Assumption 3 (SUTVA)**.**

$$Y_i = Y_i(W_i)$$

i.e. the outcome for a given observation is only affected by their own treatment assignment.

It is not obvious how to actually use (1) to estimate the CATE. We have an unknown and arbitrarily complex function $f(X_i)$ along with a treatment effect that changes based on the covariates. To proceed, we define the propensity score as

$$e(x) = \Pr[W_i = 1 \,|\, X_i = x] = \mathbb{E}\left[W_i \,\middle|\, X_i = x\right]$$

and the conditional mean of Y as

$$m(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x\right] = f(x) + \tau(x)e(x).$$

We can now rewrite (1) as

$$Y_i - m(x) = \tau(x)(W_i - e(x)) + \epsilon_i. \tag{2}$$

This formulation is typically attributed to Robinson (1988). To estimate $\tau(x)$, we assume there exists some neighborhood $\mathcal{N}(x)$ around x with some constant τ. Then for $X_i \in \mathcal{N}(x)$ the rewritten equation (2) becomes

$$Y_i - m(X_i) = \tau(W_i - e(X_i)) + \epsilon_i.$$

[4]Note that the CATE is distinct from the individual specific treatment effect, $Y_i(1) - Y_i(0)$. The CATE is still an average treatment effect, but targeted towards groups of the data that are characterized by their covariates.

Notice that if we are able to estimate the nuisance parameters $m(X_i)$ and $e(X_i)$, then we can do a simple residual-on-residual regression to estimate τ. This is a good fit for standard machine learning algorithms since we only require good predictions. One potential issue with estimation of these parameters is that naive estimation of $e(X_i)$ and $m(X_i)$ can lead to biased estimates of $\tau(x)$. Chernozhukov et al. (2018) show that if we form estimates of these nuisance parameters in a cross-fitting fashion, i.e. predicting observation i without using itself in the process, then we can obtain unbiased estimates of $\tau(x)$. A typical way to do this is to split the data into two halves, train a machine learning model on one half, and predict on the other. Our estimate of $\tau(x)$ is then

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} \mathbb{1}_{X_i \in \mathcal{N}(x)} (Y_i - \hat{m}^{-i}(X_i))(W_i - \hat{e}^{-i}(X_i))}{\sum_{i=1}^{n} \mathbb{1}_{X_i \in \mathcal{N}(x)} (W_i - \hat{e}^{-i}(X_i))^2}$$

where the $-i$ superscripts denote estimates obtained with sample splitting.

To find such neighborhoods $\mathcal{N}(x)$, we employ the causal forest algorithm. Broadly speaking, the algorithm The causal forest algorithm partitions the covariate space based on treatment effect heterogeneity and creates weights that measure how similar any observation is to a given $x \in \mathcal{X}$ in order to create "neighborhoods" of x. In more detail:

1. B trees are built from bootstrapped samples of the dataset. Each tree is grown by a splitting criterion that aims to maximize the heterogeneity of the treatment effect in child nodes[5]. Appendix C provides a more detailed explanation of how the trees are grown.

2. From this "forest" of trees, obtain weights $\alpha_i(x)$ defined as

$$\alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{1}_{X_i \in L_b(x)}}{|L_b(x)|}$$

   where $L_b(x)$ is the terminal node of tree b that x falls into. Note that this weight is the frequency that the i-th observation falls into the same terminal node as x, adjusted for the size of the terminal node. Intuitively, these weights capture how similar each observation is in treatment effect to the given x.

3. $\hat{\tau}(x)$ is then estimated by doing a weighted residual-on-residual regression with the weights $\alpha_i(x)$, giving us the estimates

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{n} \alpha_i(x)(Y_i - \hat{m}^{-i}(X_i))(W_i - \hat{e}^{-i}(X_i))}{\sum_{i=1}^{n} \alpha_i(x)(W_i - \hat{e}^{-i}(X_i))^2}.$$

   An example of the effectiveness of causal forests for estimating nonlinear treatment effect heterogeneity is provided in appendix D for the interested reader.

2.2 Empirical Strategy

The R package `grf` (Tibshirani et al., 2024) provides an easily usable implementation of the causal forest algorithm. After obtaining estimates of the CATE, it is necessary to analyze said estimates to determine if there is heterogeneity. I will first split the CATE estimates into three quantiles and provide a plot of the average treatment effect (ATE) within each quantile along with 95% confidence intervals. The ATE is estimated using an augmented inverse propensity weighted estimator (AIPW). Broadly speaking, this estimator has several desirable properties over other alternatives, even in a randomized setting. For more details, see appendix E.

One important issue with ranking the CATE estimates into quantiles is that our ranking is biased if for any two observations we were to use one of them to estimate the other. To correct for this, we will use the following procedure during the estimation process:

[5] A "node" refers to a subset of the data that results from the tree being grown. All trees begin with the "root node" i.e. the whole dataset before splitting into "child nodes."

1. Divide the dataset into K subsets.
2. Cycle through each subset and estimate CATEs using data from the other $K-1$ subsets.
3. For each subset that is being predicted, rank the estimates into quantiles.

This ensures that any comparison of the CATE estimates is done using estimates that are not dependent of each other. When there is clear heterogeneity, we should see monotonic behavior in the ATE estimates across quantiles.

Following this, I will make use of a variable importance measure provided by Bénard and Josse (2023) which computes the proportion of treatment effect variance that is lost when removing a given covariate. This acts as a heuristic for uncovering which covariates are important in determining the CATEs.

I will also fit a doubly robust best linear projection of the CATEs onto the covariates as a linear approximation of how the CATEs change with the covariates. That is, I fit the model

$$\hat{\Gamma}_i = \beta_0 + X'\beta_1 + \epsilon$$

where $\hat{\Gamma}$ is a doubly robust correction of the CATE estimates and X is the vector of covariates. See appendix E for more information on $\hat{\Gamma}$. Intuitively, this is a linear approximation of the effect that each covariate has on the treatment effect. This is another heuristic that may provide a clearer picture of which covariates are important in determining the CATEs when combined with the variance importance measure. The package `grf` provides the function `best_linear_projection` to do this.

The final analysis of heterogeneity will be to compute the ATE and 95% confidence intervals for sections of the dataset split based on covariates that were indicated as important by our previous analyses. Significant differences in the ATE estimates would confirm that a given covariate contributes to heterogeneity of the treatment effect.

3 Data

The original dataset from Oreopoulos (2011) was constructed through a field experiment conducted in the Greater Toronto Area, with additional data collection in Montreal. Between April 2008 and September 2009, a total of 12,910 fictitious resumes were sent in response to 3,225 job postings across various occupations.

These resumes were crafted to represent both recent immigrants and native workers. Importantly, among the resumes representing native workers (those with Canadian education and experience), approximately half were assigned ethnic names that could be perceived as Indian, Chinese, Pakistani, or Greek. Among the native resumes with Chinese last names, a portion was assigned an English first name. The remaining resumes varied in their combinations of ethnic names, educational backgrounds, and work experiences to represent recent immigrants from China, India, Pakistan, and Britain. Four main types of resumes were created, ranging from those with English names and all-Canadian credentials to those with completely foreign names, education, and work experience[6]. All resumes included a bachelor's degree, as this was a requirement for virtually all immigrants arriving under the Canadian immigration point system. The specific degree listed was chosen to match the job posting's occupation category. Approximately half of the resumes listed degrees from universities ranked in the Top 200 according to the 2008 QS World University Rankings, while the other half were from less prestigious institutions. Each resume was randomly assigned characteristics such as the applicant's name, educational institution, work history, and additional qualifications like language skills or extracurricular activities[7]. A full list of the names and additional qualifications can be found in the original paper.

[6] Oreopoulos constructed work experience sections using actual online resumes as templates. See the original paper for a detailed description.

[7] Extracurricular activities comprised volunteer initiatives (e.g. Habitat for Humanity), social interests (e.g. competitive sports), and proactive work skills (e.g. decision-making abilities).

The job postings generally had requirements of 3 to 7 years of experience and an undergraduate degree. Oreopoulos ignored positions that specifically required at least a graduate degree, North American experience, or certification. Typically, four resumes of increasing levels of "foreignness" were sent to each employer over a 2 to 3 day period in a random order. Responses from employers were collected via phone and email, with callbacks recorded if the employer requested contact from the applicant.

For my analysis, I restricted the dataset to include only resumes representing native Canadians (those with Canadian education and experience). This restriction was necessary due to the structure of the original experiment. In the original dataset, all resumes with either foreign experience or foreign education also have a foreign name (which includes British names). It is thus impossible to adequately separate the effect of a foreign name from the effects of foreign education or experience for methods like the causal forest, which rely on comparing similar units with different treatments. For example, there are no resumes with Indian education or experience paired with an English-Canadian name. By focusing solely on resumes with Canadian education and experience, the sample has systematic variation only through the ethnicity of names. This allows me to more accurately estimate the effect of ethnic names on callback rates for native Canadians without the confounding factors introduced by the lack of comparable foreign-experienced resumes with English names. For similar identification reasons, I also excluded observations with Chinese last names and English first names.

See appendix F for a summary of the restricted dataset, though Oreopoulos (2011) also provides more detailed tables of summary statistics.

4 Results

Figure 1 is a histogram of the CATE estimates along with the estimated ATE and its 95% confidence intervals as a heuristic visualization of the output of the causal forest. The estimated ATE of the overall dataset is approximately 0.041, meaning that the average effect of having an English name on callback rates is a 4.1% increase. This is quite close to the original estimate of Oreopoulos (2011). At first glance, it would appear as though there is some heterogeneity in the treatment effect. However, the histogram by itself cannot separate actual heterogeneity from statistical noise.

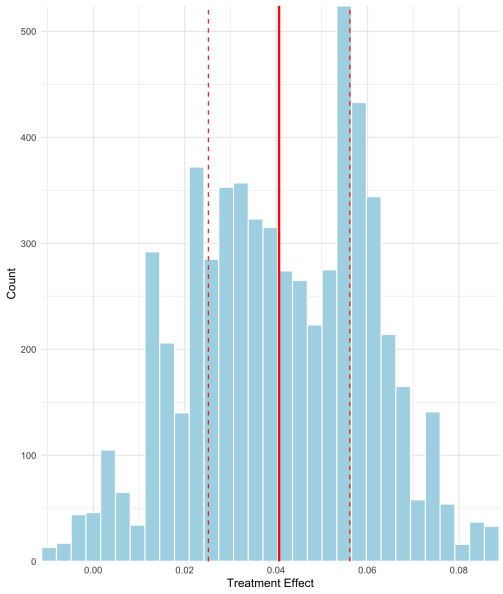

Figure 1: Histogram of CATE estimates with ATE and 95% confidence intervals

To analyze whether heterogeneity exists and what covariates are relevant, I proceed with the quantile splitting procedure described in section 2. Figure 2 shows the ATE estimates and 95% confidence intervals for the quantiles of the CATE estimates. As mentioned in section 2.2, when there is noticeable heterogeneity, we would expect monotonic behavior as the quantiles increase. This happens here, but the CATE estimates are still noisy enough that we cannot differentiate between quantile 1 and quantile 2 with reasonable confidence. We can see, however, that quantile 3 is distinct from quantile 1. In other words, the causal forest was able to distinguish individuals with high treatment effects from those with lower treatment effects.

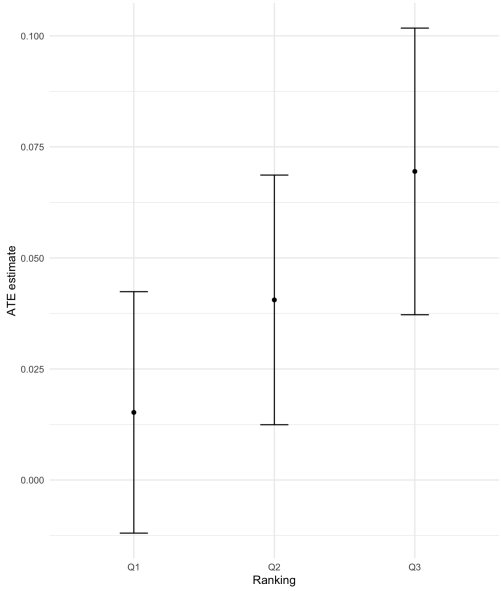

Figure 2: Quantiles of CATE Estimates

Next, I provide the results of the variable importance measure in table 1. Recall from section 2.2 that this measure computes the proportion of treatment effect variance that is lost when removing a given covariate. We can see that the most notable covariate is having a feminine name[8], while having a top 200 bachelor's, high quality work experience, and extracurricular skills listed appear to explain similar proportions of the treatment effect variation.

Variable	Importance
Master's degree	0.012
Female	0.600
Top 200 bachelor's	0.113
High quality work experience	0.126
Language skills	0.036
Extracurricular skills	0.132
Multinational work experience	0.026

Table 1: Variable Importance Results

Table 2 shows the results of fitting a doubly robust best linear projection of the CATEs onto the covariates. The coefficients cannot be interpreted as partial effects of the covariates unless the

[8]The covariate name is "Female" out of convenience, but strictly speaking the employer only knows the gender through the name.

true model were linear in covariates, but we may still glean some understanding of treatment effect heterogeneity from the projection. We see that, in agreement with the variance importance measure, candidates with a bachelor's degree from a top 200 institution, those who list extracurricular activities on their resumes, and those with feminine names exhibit distinct treatment effects compared to their counterparts. However, we can see that the BLP finds no significant effect of having high quality work experience.

Term	Estimate	Std. Error	T-Statistic	P-value
(Intercept)	0.064***	0.015	4.287	0.000
Master's degree	0.001	0.019	0.030	0.976
Female	0.040***	0.012	3.382	0.001
Top 200 bachelor's	−0.030*	0.014	−2.176	0.030
High quality work experience	−0.013	0.021	−0.597	0.551
Language skills	0.010	0.021	0.506	0.613
Extracurricular skills	−0.030***	0.009	−3.481	0.001
Multinational work experience	−0.022	0.019	−1.175	0.240

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$

Table 2: Best Linear Projection Results

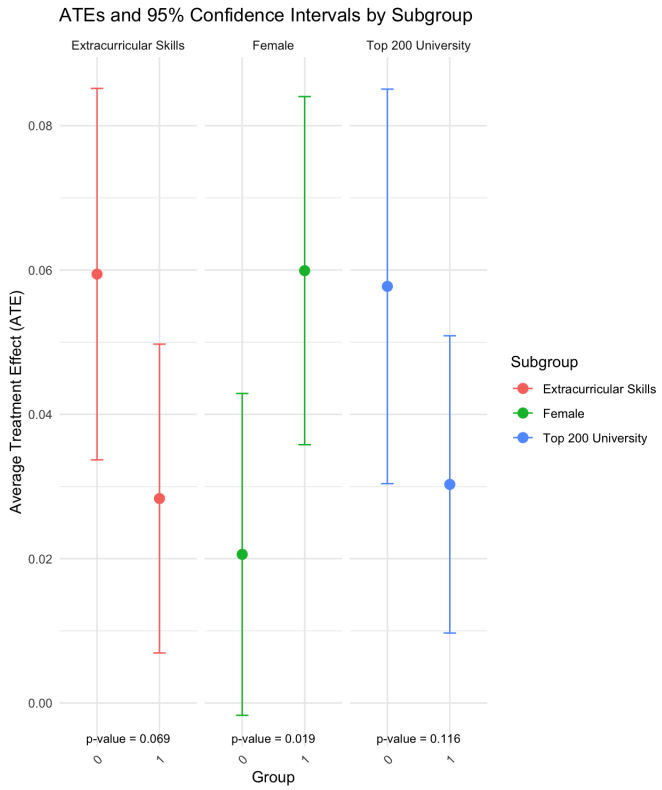

Figure 3: ATE Analysis of Important Covariates

Motivated by the variable importance measure and the BLP, I split the dataset along the covariates of having a feminine name, having a top 200 bachelor's degree, and having extracurricular skills listed

on the resume, as these are the covariates that both heuristics found to be notable. I then compute the ATE and 95% confidence intervals for each subgroup. The results are shown in figure 3. Note that due to the randomized nature of the dataset along, we may interpret the difference in ATEs as an estimate of the true effect of changing a given covariate. We see that having a feminine name distinctly increases the effect of having an english name on callback rates by about 4%, with a statistically significant difference in ATEs among those who are perceived as men versus as women. The other two covariates also appear to be notable with a negative impact on treatment effect, as expected from the BLP estimates, though they do not pass the 5% significance threshold for rejecting the null of there being no difference in the ATEs. Regardless, the p-values are low enough that we may want to take the idea of heterogenous treatment effects among these subgroups seriously.

5 Conclusion

This paper builds on the seminal work of Oreopoulos (2011) by employing the causal forest algorithm to explore possible heterogeneity in the effect of having an English-sounding name on callback rates in the Canadian labor market. I find that the impact of an English name is not uniform across all applicants. Notably, having a feminine name significantly amplifies the positive effect of having an English name on callback rates. Additionally, my findings suggest that possessing a degree from a top 200 university and listing extracurricular activities on a resume mitigate the advantage conferred by an English name, though the impact from these two covariates is less certain. These results are in contrast to the original work by Oreopoulos, which did not find any evidence of heterogeneity in the effect of having an English-sounding name.

My findings suggest that the intersectionality of ethnicity and gender plays a crucial role in hiring decisions, pointing to a more nuanced form of discrimination that existing research has not fully captured. There have been studies on the nature of gender discrimination in the work place (Verniers and Vala, 2018; Exley et al., 2020), but no clear link between gender and ethnicity has been established that explains the differential treatment uncovered in this paper. There is also scant evidence that explains why relationships between education, extracurricular activities, and the treatment effect of having an English name exist. The theory of statistical discrimination may apply here (Aigner and Cain, 1977). For instance, candidates with a degree from a top 200 university or those who list extracurricular activities might be perceived as more well-rounded or high-achieving, reducing the need for employers to rely on name ethnicity as a heuristic for competence. These attributes could signal qualities like leadership, teamwork, or cultural fit, which diminish the emphasis on having an English name as a proxy for suitability. However, this is purely speculative and further research is needed to understand the mechanisms behind these relationships.

The use of causal machine learning techniques, specifically the causal forest algorithm in this paper, represents a methodological advancement in the study of labor market discrimination. Non-parametric approaches that utilize machine learning allow for the detection of heterogeneous treatment effects by forgoing the need to impose strict functional forms, thereby uncovering patterns that traditional econometric methods might miss. The success of this approach in detecting meaningful heterogeneity where standard regression models didn't underscores its potential for future research in labor economics and other fields where treatment effects may vary across subpopulations.

References

D. J. Aigner and G. G. Cain. Statistical Theories of Discrimination in Labor Markets. *ILR Review*, 30 (2):175–187, January 1977. ISSN 0019-7939, 2162-271X. doi: 10.1177/001979397703000204. URL http://journals.sagepub.com/doi/10.1177/001979397703000204.

S. Athey and G. W. Imbens. Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, 11(1):685–725, August 2019. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev-economics-080217-053433. URL https://www.annualreviews.org/doi/10.1146/annurev-economics-080217-053433.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2), April 2019. ISSN 0090-5364. doi: 10.1214/18-AOS1709. URL https://projecteuclid.org/journals/annals-of-statistics/volume-47/issue-2/Generalized-random-forests/10.1214/18-AOS1709.full.

M. Bertrand and S. Mullainathan. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013, September 2004. ISSN 0002-8282. doi: 10.1257/0002828042002561. URL https://www.aeaweb.org/articles?id=10.1257/0002828042002561.

C. Bénard and J. Josse. Variable importance for causal forests: breaking down the heterogeneity of treatment effects, August 2023. URL http://arxiv.org/abs/2308.03369. arXiv:2308.03369 [stat].

M. Carlsson and D.-O. Rooth. Evidence of ethnic discrimination in the Swedish labor market using experimental data. *Labour Economics*, 14(4):716–729, August 2007. ISSN 0927-5371. doi: 10.1016/j.labeco.2007.05.001. URL https://www.sciencedirect.com/science/article/pii/S0927537107000358.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, February 2018. ISSN 1368-4221, 1368-423X. doi: 10.1111/ectj.12097. URL https://academic.oup.com/ectj/article/21/1/C1/5056401.

D. Clarke, J. P. Romano, and M. Wolf. The Romano-Wolf Multiple Hypothesis Correction in Stata. *SSRN Electronic Journal*, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3513687. URL https://www.ssrn.com/abstract=3513687.

C. L. Exley, M. Niederle, and L. Vesterlund. Knowing when to ask: The cost of leaning in. *Journal of Political Economy*, 128(3):816–854, 2020. doi: 10.1086/704616.

J. J. Heckman. Detecting Discrimination. *Journal of Economic Perspectives*, 12(2):101–116, June 1998. ISSN 0895-3309. doi: 10.1257/jep.12.2.101. URL https://www.aeaweb.org/articles?id=10.1257/jep.12.2.101.

S. Howard. Augmented Inverse Propensity Weighting for Randomized Experiments, March 2023. URL https://www.stevehoward.org/blog/augmented-inverse-propensity-weighting-for-randomized-experiments.html.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108 (2):299–319, May 2021. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asaa076. URL https://academic.oup.com/biomet/article/108/2/299/5911092.

P. Oreopoulos. Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Thirteen Thousand Resumes. *American Economic Journal: Economic Policy*, 3(4):148–171, November 2011. ISSN 1945-7731, 1945-774X. doi: 10.1257/pol.3.4.148. URL https://pubs.aeaweb.org/doi/10.1257/pol.3.4.148.

D. Pager. The Mark of a Criminal Record. *American Journal of Sociology*, 108(5):937–975, March 2003. ISSN 0002-9602, 1537-5390. doi: 10.1086/374403. URL http://www.journals.uchicago.edu/doi/10.1086/374403.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, September 1994. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1994.10476818. URL https://www.tandfonline.com/doi/full/10.1080/01621459.1994.10476818.

P. M. Robinson. Root-N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931, July 1988. ISSN 00129682. doi: 10.2307/1912705. URL https://www.jstor.org/stable/1912705?origin=crossref.

J. Tibshirani, S. Athey, E. Sverdrup, and S. Wager. *grf: Generalized Random Forests*, 2024. URL https://CRAN.R-project.org/package=grf. R package version 2.3.2.

C. Verniers and J. Vala. Justifying gender discrimination in the workplace: The mediating role of motherhood myths. *PLOS ONE*, 13(1):e0190657, January 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0190657. URL https://dx.plos.org/10.1371/journal.pone.0190657.

A Robustness Check

The Romano-Wolf procedure uses resampling procedures to control the familywise error rate. In particular, Romano-Wolf provides more power, meaning the ability to correctly reject null hypotheses, compared to other corrections like the Bonferroni correction.

As a short overview, the Romano-Wolf procedure begins by calculating the original test statistics for each hypothesis. Then, a large number of bootstrap samples are generated from the original dataset and test statistics are computed for each sample. These bootstrap test statistics are centered by subtracting the original test statistics to reflect the null hypothesis. For each bootstrap sample, the maximum test statistic (in absolute value) is identified to create a distribution of maximum values. The adjusted p-values are calculated by comparing the original test statistics to this distribution. More specific details can be found in Clarke et al. (2020).

Covariate	Original p-value	Adjusted p-value
Female	0.019	0.055
Top 200 bachelor's	0.116	0.130
Extracurricular skills	0.069	0.130

Table 3: Original and Adjusted p-values for Testing Difference in ATEs

Table 3 shows the original and adjusted p-values for the testing of differences in ATEs among subgroup splits. The adjusted p-values are pushed out of conventional significance levels, but not by much. The interpretation of an adjusted p-value is also slightly different than normal. Specifically, the adjusted p-value gives the significance level such that if all hypotheses with lower p-values than the adjusted were rejected, then the familywise error rate would be controlled at the adjusted p-value. Intuitively, this means we can be "87% confident" that the differences in ATEs are not false positives.

B Identification of the CATE

To see why the assumptions discussed in section 2.1 are necessary for identification of the CATE, notice the following:

$$
\begin{aligned}
\tau(x) &\coloneqq \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right] \\
&= \mathbb{E}\left[Y_i(1) \,\middle|\, X_i = x\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, X_i = x\right] \\
&= \mathbb{E}\left[Y_i(1) \,\middle|\, W_i = 1, X_i = x\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, W_i = 0, X_i = x\right] && \text{(Unconfoundedness)} \\
&= \mathbb{E}\left[Y_i \,\middle|\, W_i = 1, X_i = x\right] - \mathbb{E}\left[Y_i \,\middle|\, W_i = 0, X_i = x\right] && \text{(SUTVA).}
\end{aligned}
$$

Overlap is necessary due to the last equation – for a given covariate value x, there must be some observations with $W_i = 1$ and some with $W_i = 0$.

C Growing Trees

Individual trees are grown in the following manner:

1. A random subsample is drawn from the full dataset. A single parent node is created from this sample.
2. The node is split into two child nodes, which are then split recursively. That is, each child node becomes a parent node that is then split again until there are no more nodes to split.
3. In deciding how to split a node, the algorithm proceeds as follows:

(a) A random subset of variables are chosen from the covariates as candidates to split on.

(b) For each variable, all possible values of that variable are considered as potential split points. The best split is chosen by maximizing the criterion

$$n_L n_R (\hat{\tau}_L - \hat{\tau}_R)^2$$

where n_L and n_R are the number of observations in the left and right child nodes, and $\hat{\tau}_L$, and $\hat{\tau}_R$, are estimates of the treatment effects in the left and right child nodes that are obtained through residual-on-residual regression for each possible split point. A formal justification of this criterion is given in Athey et al. (2019), though intuitively the criterion aims to maximize the heterogeneity of the treatment effect in the child nodes.

D Example of Causal Forest Performance

Figure 4 is an example of the performance of a causal forest compared to linear regression in estimating nonlinear treatment effect where the true data generating process involved 20 covariates, of which only X_1 and X_3 influenced the effect of treatment.

The linear regression model used standard interaction terms of treatment with X_1 and X_3. Notice how the linear regression model is unable to capture the nonlinear relationship between the treatment effect and the covariates, while the causal forest is able to much more accurately estimate the treatment effect. Theoretically, linear regression could do better if the researcher specified a more complex model, but this would require some knowledge of the true data generating process.

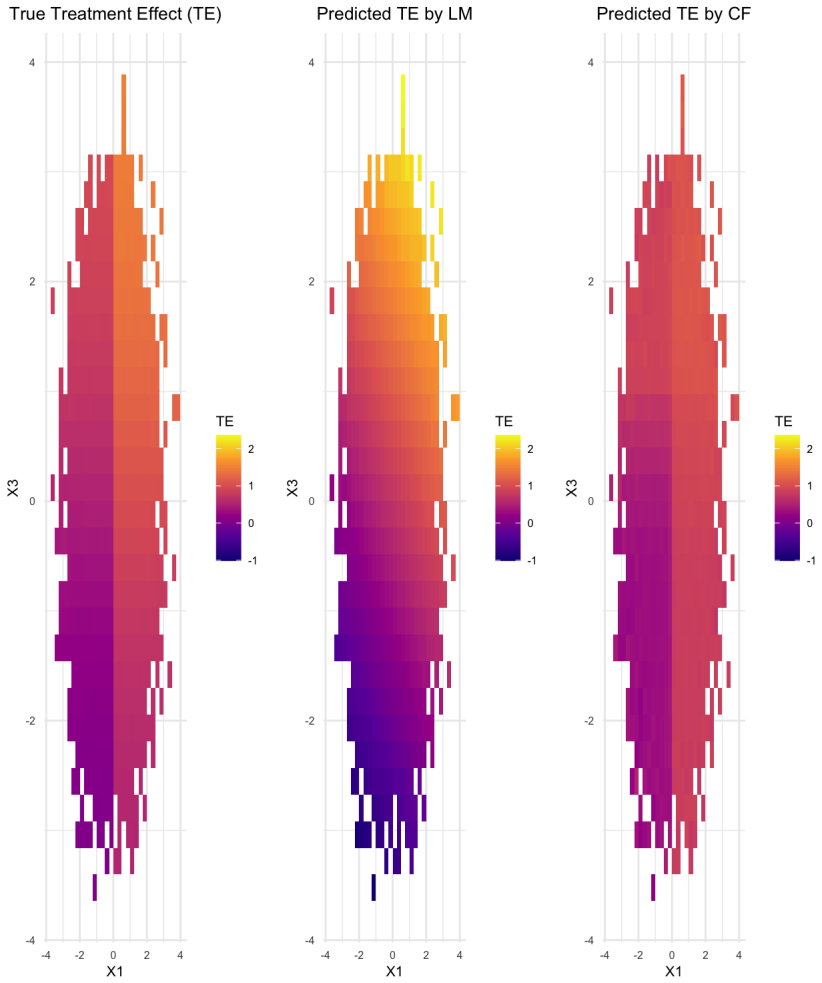

Figure 4: Performance of Causal Forest and Linear Regression in Estimating Treatment Effect

E Estimating the ATE

To estimate the ATE, first define the estimated doubly robust score

$$\hat{\Gamma}_i = \hat{\tau}(X_i) + \frac{W_i - \hat{e}^{-i}(X_i)}{\hat{e}^{-i}(X_i)[1 - \hat{e}^{-i}(X_i)]}[Y_i - \hat{\mu}^{-i}(X_i, W_i)] \tag{3}$$

where $\hat{\mu}^{-i}(X_i, W_i)$ is an estimate of the realized conditional mean $\mathbb{E}\left[Y_i \,\middle|\, X_i = x, W_i = w\right]$ for observation i. The ATE estimate is then the average of these doubly robust scores. This is known as the augmented inverse propensity weighted estimator (AIPW), generally attributed to Robins et al. (1994):

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\tau}(X_i) + \frac{W_i - \hat{e}^{-i}(X_i)}{\hat{e}^{-i}(X_i)[1 - \hat{e}^{-i}(X_i)]}[Y_i - \hat{\mu}^{-i}(X_i, W_i)]\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{\Gamma}_i$$

There are two main advantages to this estimator. First, it is "doubly robust" in the sense that as long as either $\hat{\mu}$ or $\hat{e}(x)$ is correctly specified, the estimator will be consistent. Second, if the estimate of $\hat{\mu}$ is close to the true value, then this estimator will be efficient among a large class of estimators, including a standard difference-in-means or regression in a randomized setting. The variance properties of averaging 3 are the main reason why the best linear projector does not simply use the CATE estimates as the outcome variable. Howard (2023) provides a good overview of the AIPW estimator.

F Table of Summary Statistics

Characteristics of resume	Name Type	
	English (1)	Foreign (2)
Female	0.51	0.51
Top 200 world ranking university	0.59	0.66
Extra curricular activities listed	0.60	0.61
Fluent in French and other languages	0.25	0.26
Canadian master's degree	0.21	0.20
Multinational firm work experience	0.28	0.21
High quality work experience	0.33	0.30
Name ethnicity		
English-Canadian	1.00	0.00
Indian	0.00	0.39
Pakistani	0.00	0.16
Chinese	0.00	0.33
Greek	0.00	0.12
Number of Observations	3026	2997

Table 4: Proportion of Resumes Sent with Particular Characteristics by Name Type