CERGE-EI

Center for Economic Research and Graduate Education - Economics Institute

Charles University

Causal Machine Learning for Heterogeneous Treatment Effects

An Application on Optimal Treatment Assignment

Master's thesis

Author: Bc. Klaus Hajdaraj

Supervisor: Paolo Zacchia, Ph.D.

Year of defense: 2025

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, January 1, 2025

Klaus Hajdaraj

Abstract

With the rising popularity of machine learning for uncovering complex patterns, there is growing interest in leveraging these techniques to understand how interventions affect individuals differently based on their characteristics, a concept known as heterogeneity (HTE). This paper compares two machine learning methods for predicting HTEs for optimal treatment assignment or so-called targeting: the causal forest (CF), a direct tree-based method, and the causal neural network (CNN), an indirect deep learning method. I use an empirical dataset from an online experiment on incentivizing manual labour to compare the methods. I show that CF outperforms CNN; assigning individual optimal treatments based on CF yields higher outcomes than assigning the overall best treatment to all individuals. Further, I address the winner's curse in the optimal targeting context by introducing two shrinkage techniques: the James-Stein and the Variance shrinkers, which improve the performance of ML methods in assigning the optimal treatments. This study contributes to the literature by providing a detailed guideline for selecting and comparing ML methods for optimal targeting and introducing shrinkage techniques to adjust upward bias (overestimation). The findings highlight the importance of accurate HTE estimation in improving optimal targeting, and recommend development of personalized treatments. Personalized treatments can improve overall outcomes by tailoring policies to individuals' characteristics.

JEL Classification	C14, C21, C45, C52, C53
Keywords	optimal targeting, heterogeneity, machine learning, causal forest, neural networks, shrinkage estimator
Title	Causal Machine Learning for Heterogeneous Treatment Effects: An Application on Optimal Treatment Assignment
Author's e-mail	Klaus.Hajdaraj@cerge-ei.cz
Supervisor's e-mail	Paolo.Zacchia@cerge-ei.cz

Abstrakt

S rostoucí popularitou použití strojového učení při odhalování komplexních souvislostí roste zájem o využití těchto technik k pochopení toho, jak intervence ovlivňují jednotlivce odlišně v závislosti na jejich charakteristikách, což je koncept známý jako heterogenita (HTE). Tato práce porovnává dvě metody strojového učení pro předpovídání HTE za účelem optimálního přiřazení treatmentu nebo tzv. cílení: kauzální les (CF), přímou metodu založenou na stromech, a kauzální neuronovou síť (CNN), nepřímou metodu hlubokého učení. K porovnání metod používám empirický soubor dat z online experimentu zaměřeného na motivaci k manuální práci. Ukazuji, že CF překonává CNN; přiřazení individuálních optimálních treatmentů na základě CF přináší lepší výsledky než přiřazení celkového nejlepšího ošetření všem jednotlivcům. Dále se zabývám prokletím vítěze v kontextu optimálního cílení zavedením dvou technik smršťování: smršťovač Jamese-Steina a smršťovač rozptylu, které zlepšují výkonnost ML metod při přiřazování optimálních treatmentů. Tato studie přispívá k literatuře tím, že poskytuje podrobný návod pro výběr a porovnání ML metod pro optimální cílení a zavádí techniky smršťování pro úpravu vychylení směrem nahoru (nadhodnocení). Zjištění zdůrazňují význam přesného odhadu HTE pro zlepšení optimálního cílení a doporučují vývoj personalizovaných treatmentů. Personalizované cílení může zlepšit celkové výsledky díky přizpůsobení zásahů charakteristikám jednotlivců.

Klasifikace JEL	C14, C21, C45, C52, C53
Klíèová slova	optimální cílení, heterogenita, strojové učení, kauzální les, neuronové sítě, smršťovací estimátor
Název práce	Kauzální strojové učení a heterogenní efekty treatmentu: Aplikace na optimální přiřazení treatmentu
E-mail autora	`Klaus.Hajdaraj@cerge-ei.cz`
E-mail vedoucího práce	`Paolo.Zacchia@cerge-ei.cz`

Acknowledgments

I would like to express my gratitude to my supervisor, Paolo Zacchia, Ph.D., for his honest feedback and invaluable guidance throughout this research journey. I extend my sincere thanks to the ASC for their guidance in helping me become a more effective writer. Finally, I owe lots of thanks to my family, friends, and colleagues, for their constant support.

Bibliographic Record

Hajdaraj, Klaus: *Causal Machine Learning for Heterogeneous Treatment Effects: An Application on Optimal Treatment Assignment.* Master's thesis. CERGE-EI, Center for Economic Research and Graduate Education - Economics Institute, Charles University, Prague. 2025, pages 76. Advisor: Paolo Zacchia, Ph.D.

Contents

[1]

[1]Previous versions of this paper were submitted in *Research Writing 1*, *Research Writing 2* and *Microeconometrics 2* during Spring 2023, Fall 2023, and Spring 2024 at CERGE-EI.

Introduction

Economists and professionals across various fields are keenly interested in understanding the causal effects of policies and interventions. Over the past few decades, this interest has spurred significant advancements in microeconometrics and statistics, particularly in identifying and estimating different average causal effects (see, e.g., Imbens and Wooldridge 2009; Athey and Imbens 2017, and references therein). However, focusing solely on average effects often overlooks how these causal effects vary across individuals with different observable characteristics. Therefore, assuming that everyone benefits equally from the intervention is often unrealistic. Intuitively, treatment effects can vary across different subgroups within the population, depending on their observable characteristics or covariate values. In other words, there is heterogeneity in the treatment effects (HTE). For instance, based on individual observed characteristics, identifying which individuals benefit most from active labour market policies, promotional campaigns, or medical treatments is crucial for efficient allocation of public and private resources.

Recent progress in machine learning (ML) techniques, coupled with the availability of large datasets and developments in causal inference literature, has brought estimation of HTEs to the forefront of research. Central to this discourse is whether to use a particular treatment and whether we can determine which treatment will be optimal. Consequently, a primary application of identifying HTEs is in the assignment of optimal treatments, where the effectiveness for each individual is assessed based on their predicted treatment effect.

Various research disciplines have developed systematic methods for estimating causal HTEs. These methods adapt standard ML algorithms to flexibly estimate heterogeneity across potentially large numbers of covariates. Available estimators employ techniques such as random forests (Wager and Athey 2018; Athey et al. 2019), LASSO (Tian et al. 2014; Chen et al. 2017), deep neural networks (Johansson et al. 2016; Schwab et al. 2018), and Bayesian ML approaches (Taddy et al. 2016).

Applied studies utilizing these methods have recently emerged in economics (Andini et al. 2018; Ascarza 2018; Strittmatter 2023). In particular, the medical sector and the field of health economics have shown increasing interest in patient-centred outcomes (Willke et al. 2012), which are central to HTE estimation and optimal treatment assignment. Other economic applications, such as incentivizing effort in employee management (DellaVigna and Pope 2018) or evaluating the effectiveness of public policies (e.g., education, tax policy) should also be considered. From a business standpoint, understanding the varying impacts of advertising or marketing offers on consumer purchases has also become of significant interest.

Given the rapidly expanding literature on ML methods for treatment effect estimation, I investigate a pertinent question about which methods perform best for optimal treatment assignment in a setting in which **multiple treatments are available** and how they can effectively be compared using empirical data.

In this paper, I present the causal forest (CF) method (Wager and Athey 2018), a tree-based approach for *direct* estimation of HTEs, and causal neural networks (CNN) (Farrell et al. 2021), an *indirect* semiparametric method based on feed-forward neural networks. I compare these methods for predicting HTEs for optimal treatment assignment using an empirical dataset on incentivizing manual labour in an online experiment, employing a comparison technique based on Hitsch et al. (2024) and the *R-learner* loss function structure from Nie and Wager (2021).

To compare the two methods, I use data [2] from an experiment conducted by Opitz et al. (2024) on Amazon Mechanical Turk (MTurk), a platform primarily used for small-scale contract labour but increasingly popular for behavioural experiments. Using MTurk not only facilitates collection of large-scale samples and reduces payoff costs but also provides a more diverse participant pool than do traditional university-based experiments, which potentially offers better external generalization (Follmer et al. 2017).

I find that the CF method achieves the best performance in assigning individuals the optimal treatment from a given set of treatments. The results indicate that selecting treatments based on the highest predicted treatment effect from the CF results in higher outcome levels than simply assigning the overall best-performing treatment to all individuals. In contrast, the CNN method performs only marginally better than random treatment assignment and significantly worse than the CF method.

Moreover, a significant challenge in optimal treatment assignment using HTEs is the *winner's curse*. Treatments with overestimated effects are more likely to be identified as optimal because the selection is based on the highest predicted treatment effect. Therefore, I investigate whether shrinkage estimators can enhance the performance of ML methods in assigning optimal treatments and how their application affects the initial results compared to when shrinkage techniques are not used. I outline the concept of the winner's curse, propose shrinkage estimators as a potential solution, and evaluate their effectiveness when applied to the ML predictions on the empirical dataset. In addition, I introduce a modified version of both shrinkage methods that adjusts estimates toward the average treatment effect across all considered treatments instead.

I find that employing shrinkage methods can enhance the performance of predictions in most cases. The shrinkage estimators performed differently across the two models and the treatment subsets. Notably, shrinkers that adjust predictions toward the overall av-

[2] Thesis replication files, including Python source code and data are available at: https://github.com/klaushajdaraj/ml-treatment-effects

erage outcome performed less accurately when analyzing all six treatments but showed improved performance when focused on a subset of four similar treatments. Overall, the James-Stein Shrinker resulted in greater performance improvements than did the Variance Shrinker. The CF method benefited significantly from applying shrinkage estimators, whereas the CNN method did not show notable improvements.

Two studies closely related to this paper are those by Opitz et al. (2024) and Hitsch et al. (2024). Opitz et al. (2024) examine the performance of targeted assignment of incentive schemes, conducting two large-scale experiments, each involving an extensive personality trait survey followed by a manual labour task. The first experiment served for pre-analysis and model training, while the second compared the treatment assignment performance using the Virtual Twin Random Forest technique. The data utilized in the empirical section of this paper is retrieved from the first experiment of Opitz et al. (2024). They show that previously captured personality traits could predict participants' work performance. Thus, employers can leverage information about worker heterogeneity to enhance the effectiveness of incentives through targeted assignments based on individual characteristics. Furthermore, Opitz et al. (2024) found that assignment based on predictions using the Virtual Twin Random Forest yielded significantly higher outcomes than assigning the overall best-performing treatment from the first experiment.

Examining whether personality traits affect participant performance is beyond the scope of my study. However, this thesis differs from Opitz et al. (2024) in two significant ways. First, I aim to empirically and thoroughly compare two ML methods for optimal treatment assignment. To provide external validity for the predictions, I employ 100 repetitions of three-fold cross-validation. Second, I address a critical issue in optimal targeting not considered by Opitz et al. (2024): the *winner's curse*. I present the problem of the winner's curse and propose two families of shrinkage estimators as a solution. The application of shrinkage estimators in the context of estimating treatment effects for optimal targeting is novel and not extensively developed in prior literature.

Additionally, although I employ a methodology similar to that used by Hitsch et al. (2024) to compare the two ML methods, this thesis differs from their study in several key aspects. First, their application focuses on customer targeting frameworks, such as companies using catalogues, emails, and display ads, whereas my empirical application centres on incentive schemes for worker performance. Second, they concentrate on Treatment Effect Projection (TEP) and causal KNN regression techniques, while I focus solely on CF and CNN.

The contributions of this paper to the growing literature on ML for optimal treatment assignment are twofold. First, I aim to provide a guideline for systematically choosing and comparing different ML estimation methods to predict optimal targeting

policies, to effectively estimate individual-level effects of targeting efforts. Second, I introduce a novel application of shrinkage estimators for ML methods in the context of optimal treatment assignment. This is the first study to use shrinkage techniques to estimate treatment effects for optimal treatment targeting.

Investigating methods for optimal treatment assignment is crucial, as it enables targeted interventions that maximize treatment efficacy while minimizing potential adverse effects. By employing ML for causal inference, researchers can more accurately predict individual responses to different treatments, and ensure that resources are assigned to those who will benefit the most. Consequently, optimal targeting can improve policymaking by offering personalized interventions.

The remainder of this paper is organized as follows: First, I discuss related literature in Section 2. Then, I present machine learning methods in Section 3. In Sections 4 and 5, I outline the data and empirical strategy, including ML training and tuning, model comparison strategy and shrinkage estimators. Further, section 6 presents the results for non-shrinkage and shrinkage use cases, while section 7 discusses the results and potential limitations. Finally, Section 8 presents the conclusions. [3]

[3]Parts of this paper were edited using *Grammarly* AI editing tool to refine the language and enhance clarity.

Literature Review

2.1 Causal Inference Fundamentals

A vast amount of econometrics literature focuses on estimating causal effects. In the last three decades, the potential outcome framework, often termed the Rubin Causal model, has become a predominant method for addressing causal inference issues. In this model, each individual i has a potential outcome $Y_i(T)$ for each treatment level T, reflecting the outcome if the individual were subject to that treatment. According to foundational work by Neyman (1923), the causal effect of a treatment or intervention is typically defined as the difference between the actual observed outcome and the hypothetical outcome that would have occurred if the treatment had not taken place. This definition essentially establishes a basis for the concept of counterfactuals, which Rubin (1974) later formalized into the extensive potential outcomes framework. This framework has since become a cornerstone in the field of causal inference research. However, while we can observe the treatment an individual receives and its resulting outcome, the outcomes for alternative treatments that the individual did not receive remain ultimately unobserved, presenting what Holland (1986) describes as the "fundamental problem of causal inference". Outside of science fiction, where parallel universes might be imagined as observable, measuring causal effects at the individual level is impossible. Consequently, researchers focus on estimating average causal effects (ATE). When applying a binary treatment T, the potential outcomes for the treatment and control group are recorded as Y^0 and Y^1, correspondingly. The treatment effect for the unit i is expressed as:

$$\text{TE} = Y_i^1 - Y_i^0$$

In this case, the ATE for the sample is calculated as:

$$\text{ATE} = E[Y_i^1 - Y_i^0]$$

Another widely studied method in the context of ATE is the average treatment effect on the treated (ATET), expressed as:

$$E[Y_i^1 - Y_i^0 | T_i = 1] = E[Y_i^1 | T_i = 1] - E[Y_i^0 | T = 1]$$

This expression illustrates the counterfactual nature of causal effects. The first term is the potentially observable metric, that represents the average outcome of the treated group. In contrast, the second term denotes the average outcome of the treated group if they had not undergone the treatment, an unobservable quantity. Thus, we can employ

an econometric **identification strategy** (Angrist and Krueger 1999) that provides a consistent estimate of this unobservable term.

2.1.1 Challenges and Methods for Estimating Causal Effects

Although randomized control trials (RCTs) are considered the gold standard for studying causal relationships, as randomization eliminates much of the bias inherent with other study designs, RCTs have drawbacks, including ethical risks, high costs in terms of time and money and problems with generalizability. Observational studies have several advantages over RCTs, including lower costs and longer timeliness. However, in many observational studies that assess the impact of a policy or intervention, estimating the average treatment effect (ATE) is a key challenge due to the presence of **selection bias** (or omitted variable bias).

Selection bias occurs because the treatment group differs from the control group for reasons beyond the treatment status per se. Merely comparing treated and untreated units using ATE may yield a misleading estimate of the causal effect. Because the issue of omitted variables is not directly related to the sampling variance but rather concerns population quantities, the difference in outcomes by observed treatment status is expressed as:

$$\begin{aligned} E[Y_i|T_i=1] - E[Y_i|T_i=0] &= E[Y_i^1|T_i=1] - E[Y_i^0|T_i=0] \\ &= \underbrace{E[Y_i^1 - Y_i^0|T_i=1]}_{\text{ATET}} + \underbrace{E[Y_i^0|T_i=1] - E[Y_i^0|T_i=0]}_{\text{selection bias}} \end{aligned}$$

The issue of selection bias drives the need for implementation of random assignment in experiments to allow estimation of treatment effects. With the random assignment of treatment T_i, $E[Y_i|T_i=1] - E[Y_i|T_i=1] = E[Y_i^1 - Y_i^0] = E[Y_i^1] - E[Y_i^0]$. Replacing $E[Y_i|T_i=1]$ and $E[Y_i|W_i=0]$ with their respective sample counterparts provides a consistent estimate of ATE.

To address potential selection bias that could affect the estimation of ATEs, policymakers must disentangle the impact of the intervention from other confounding factors influencing the outcomes. As a result, evaluating ATEs in observational studies necessitates adjustments for differences in baseline covariates, because the treatment and control groups may be imbalanced in terms of both measured and unmeasured covariates. The literature offers several approaches on how to handle missing counterfactuals in treatment effects evaluation theory.

Without making adjustments for baseline covariates, the ATE in observational studies can be estimated using a complete-case estimator expressed as follows:

$$\widehat{ATE} = \frac{\sum_1^n T_i Y_i}{\sum_{i=1}^n T_i} - \frac{\sum_{i=1}^n (1 - T_i) Y_i}{\sum_{i=1}^n (1 - T_i)} \tag{1}$$

The $\widehat{ATE}$ estimate is subject to bias and must be adjusted to account for variations in baseline covariates.

A key assumption in causal inference is the **unconfoundedness** or conditional independence assumption (Rosenbaum and Rubin 1983). This assumption asserts that, given a set of observable covariates X_i that is unaffected by the treatment, the potential outcomes Y_i^1 and Y_i^0 are independent of the treatment assignment T_i. The unconfoundedness assumption is expressed as:

$$(Y_i^1, Y_i^0) \perp T_i | X_i \tag{2}$$

Unconfoundedness goes by other names in different fields, including ignorability (Rosenbaum and Rubin 1983), the back-door criterion (Pearl 2009), and exogeneity (Wooldridge 2015).

Another crucial assumption is the **common support** condition which is expressed as:

$$0 < P(T_i = 1 | X_i) < 1, \forall i \tag{3}$$

This assumption requires that, within the observed data, every individual or unit has a non-zero probability of receiving any level of treatment. In other words, the treatment assignment must vary among the study population. This condition ensures that individuals with diverse characteristics are represented in both the treatment and control groups. Violating the positivity assumption—where certain groups have no chance of receiving the treatment—can lead to biased estimates, and create challenges in generalizing the study findings to a broader population.

Lastly, an additional key assumption is **the stable unit treatment value assumption** (SUTVA). This assumption asserts that the treatment of one sample unit does not affect the outcome of another sample unit.

Using unconfoundedness and common support assumptions, Rosenbaum and Rubin (1983) introduce the concept of the propensity score (PS), which represents the probability of receiving treatment given a set of covariates. Under the conditional independence assumption, the PS is given by

$$p(X_i) = P(T_i = 1 | X_i) = P(T_i = 1 | Y_i^1, Y_i^0, X_i) \tag{4}$$

Following that, Kreif et al. (2013) describe a widely used parametric model for estimating the propensity score, the logistic regression model, which is expressed as

$$p(X_i, \eta) = \frac{e^{\eta^T X_i}}{1 + e^{\eta^T X_i}} \tag{5}$$

where η represents a vector of parameters that can be estimated using the maximum likelihood estimate $\hat{\eta}$ based on the observational data (T_i, X_i), where $i = 1, ..., n$.

The propensity score enables two distinct methods for estimating ATEs: **inverse propensity weighting** (IPW) and **propensity score matching** (PSM). Both approaches address the differences between treatment and control groups by adjusting for baseline covariates, thereby reducing selection bias. The IPW method assigns weights to each observed outcome based on the inverse probability of it being observed, while the PSM method creates a matching control group that closely resembles the treatment group in terms of observed covariates using the propensity score. Kreif et al. (2013) show that the IPW estimate of the ATE is derived by reweighting the observed outcomes for both treatment and control groups using the inverse of the estimated probability of receiving the observed treatment:

$$\text{IPW} = \frac{1}{n}\sum_{i=1}^{n} T_i w_i Y_i - \frac{1}{n}\sum_{i=1}^{n}(1 - T_i) w_i Y_i \tag{6}$$

where w_i is expressed as $w_i = \frac{1}{p(X_i, \hat{\eta})}$. The validity of the IPW estimate depends on the proper specification of the propensity score model $p(X_i, \eta)$.

Kreif et al. (2013) suggest also using generalized linear models, $g_1(X_i, \beta_1)$ and $g_0(X_i, \beta_0)$ to estimate the treatment effects:

$$\widehat{REG} = \frac{1}{n}\sum_{i=1}^{n} g_1(X_i, \hat{\beta}_1) - \frac{1}{n}\sum_{i=1}^{n} g_0(X_i, \hat{\beta}_0) \tag{7}$$

Here, $\hat{\beta}_1$ and $\hat{\beta}_0$ represent the maximum likelihood or least squares estimates of parameter vector β_1 and β_0.

By combining the models for the propensity score, Robins et al. (1994) propose a hybrid model of the propensity score and regression methods to estimate ATE, called **augmented inverse propensity weighting** (AIPW) method, expressed as:

$$\begin{aligned}\widehat{AIPW} = \frac{1}{n}\sum_{i=1}^{n} T_i w_i [Y_i - g_1(X_i, \hat{\beta}_1)] - \frac{1}{n}\sum_{i=1}^{n}(1 - T_i) w_i [Y_i - g_0(X_i - \hat{\beta}_0)] \\ + \frac{1}{n}\sum_{i=1}^{n} g_1(X_i, \hat{\beta}_1) - \frac{1}{n}\sum_{i=1}^{n} g_0(X_i, \hat{\beta}_0)\end{aligned}$$

The validity of the AIPW estimate depends on the correct specification of either the propensity score or the outcome regression models, but not necessarily both, making

the AIPW estimate **doubly robust** (DR). Additionally, the AIPW estimate achieves semiparametric efficiency when both the propensity score and potential outcomes are correctly specified.

For more detailed information on double robust methods, see Chernozhukov et al. (2018).

2.1.2 Heterogeneity and Conditional Average Treatment Effects

Thus far, this paper has assumed that the treatment effect is uniform across all individuals in the sample, which explains the constant ATE in the previous equations. However, assuming that everyone benefits equally from the intervention is often unrealistic. Therefore, I now introduce heterogeneous treatment effects (HTE) and conditional average treatment effects (CATE) because the core focus of this thesis is identifying HTE using machine learning methods. CATE explicitly represents the difference in the expected outcome between two groups of individuals who share similar observed characteristics or covariates, and differ only in their treatment status. The "conditional" aspect of CATE refers to the fact that the treatment effect can vary across different subgroups within the population, depending on their covariate values.

To illustrate the concept of HTE, Figure 1 presents an example of a homogeneous treatment effect (Figure 1a) and an HTE (Figure 1b) within the sample. In the homogeneous case, while individual outcomes vary within each treatment group and between all treatment groups, the treatment effect $\tau(x)$ remains consistent for all individuals and matches the average treatment effect (Figure 1a). In contrast, HTE (Figure 1b) demonstrates substantial variation in how individuals respond to treatment, with some showing greater impact, others showing less impact, and some showing no impact at all. Therefore, the ATE offers limited insight at the individual level, and HTE analysis is necessary to understand how treatment effects differ across the entire sample.

To express CATE mathematically using the same notations as before, say that the data consists of observations $(Y_i, T_i, X_{1i}, ..., X_{Ji})$, $i = 1, ..., n$ with n being the number of observations and J being the number of covariates observed. As before, T_i is the individual treatment status, and $X_{1i}, ..., X_{Ji}$ denotes the J covariates for individual i.

The CATE, which represents the heterogeneous treatment effect, is defined as:

$$\text{CATE}: \ \tau(x) = E[Y_i^1 - Y_i^0 | X = x] \tag{8}$$

To effectively recover treatment effects $\tau(x)$ in the presence of heterogeneity, it is crucial to specify a comprehensive model to capture all potential subgroups where the effect remains constant. Capturing this information becomes particularly challenging with linear models. However, if the outcome equation is assumed to be linear, it is

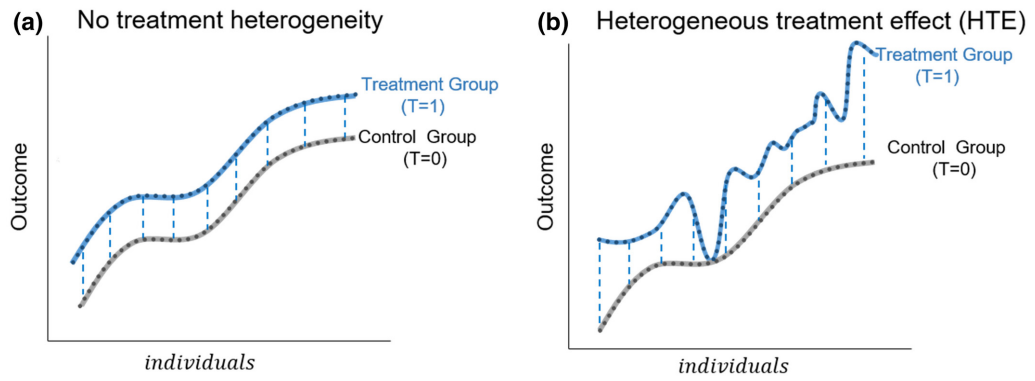

Figure 1: **Source**: Gong et al. (2021)
(a) Homogeneous treatment effect (no heterogeneity): While the treatment outcomes vary across individuals and between the treatment groups, the treatment effect (represented by the difference in outcomes between the two curves, shown by the dotted lines) remains identical for all individuals. (b) Heterogeneous treatment effect (HTE): The treatment effect differs among individuals, with some gaining more, others less, and some potentially receiving no benefit from the treatment at all.

essential to include all possible interactions among variables to minimize bias in the CATE. The main task is to identify subgroups where the treatment effect is relatively stable across all units, ensuring that the variation in covariates X_i is sufficient to define these subgroups.

Variance plays a key role because it naturally reflects heterogeneity, as indicated by the equation $\text{Var}(\tau_i) = \text{Var}(\tau(X_i)) + \text{Var}(\varepsilon_i)$. Assuming zero covariance between the terms on the right side due to unconfoundedness, an optimal approximation of τ can be achieved by maximizing $\text{Var}(\tau(X_i))$.

If a model lacks all possible interactions, it fails to identify these subgroups, resulting in a biased estimation of $\tau(x)$. Even when all relevant covariates X_i are available to ensure unconfoundedness, the bias persists unless the variance of the error term ε_i is minimized. This essentially requires full specification of the functional form of $\tau(x)$ to maximize the explained variance.

This insight is fundamental for incorporating machine learning algorithms into the estimation of heterogeneous treatment effects. This concept is revisited in the section 3 of this thesis.

2.2 Background to Optimal Treatment Assignment

2.2.1 The Optimal Treatment Assignment Problem

In this section, I formalize the problem of optimal treatment assignment. The treatment-assignment problem arises in scenarios in which a policymaker aims to maximize the

overall causal impact of interventions (treatment) on a specific outcome. Each alternative represents a distinct treatment, with the objective being to assign individuals to the treatment that maximizes their outcome. Hence, a crucial goal of empirical research on treatment effects is to equip policymakers with information that helps them to select appropriate treatments. On the other hand, automated decision-making systems frequently aim not only to predict outcomes but to actively improve them. For instance, in educational settings, a system may allocate personalized learning resources to students to boost their performance rather than simply predicting their academic success. This task broadly represents a treatment-assignment problem (Manski 2004), where each potential action aligns with a specific "treatment" (e.g., "assign extra practice" versus "no additional practice"). Ideally, each student receives the treatment associated with the most advantageous result (e.g., the one that leads to the most significant improvement in performance).

In this paper, I focus on scenarios in which decisions are independent, and the treatment-assignment policy is derived from historical (observational) data based on prior random decisions. This setup ensures that each decision impacts only one unit (or instance) and that the data remains free from selection bias. Moreover, the *stable unit treatment value* (**SUTVA**) and the *unconfoundedness* assumptions hold when a carefully designed randomized A/B test is used to gather data.

T represents the treatment-assignment variable and Y denotes the observed outcome. Within the potential outcomes framework (Rubin 1974), $Y(j)$ is defined as the outcome observed if treatment j (among k possible treatment options) is assigned, meaning $Y = Y(j)$ when $T = j$. Thus, the treatment assignment that would yield the highest average outcome is:

$$a^* = \underset{j}{\operatorname{argmax}}\, \mathbb{E}[Y(j)] \tag{9}$$

which can be be estimated by calculating the sample mean for each treatment:

$$\hat{a} = \underset{j}{\operatorname{argmax}}\, \hat{\mathbb{E}}[Y|T = j] \tag{10}$$

Equation 10 outlines a typical A/B testing method for comparing multiple treatments within a specified population. However, this method does not account for treatment assignments tailored to individuals.

Because individuals with different characteristics often differ in their responses to treatment (heterogeneous treatment effects), statistical modelling enables us to estimate the treatment-assignment policies from observational data. These tools can guide policymakers in matching individuals to the most suitable treatment based on their

personal characteristics (such as preferences and behaviours). Therefore, treatment assignment policies that distribute treatments according to individuals' observed characteristics can substantially impact outcomes.

In this paper, I determine the optimal treatments for individuals or for specific subpopulations based on heterogeneity on individual characteristics. Assuming individuals differ based on a set of variables (features) X, we can interpret a feature vector x as representing a subpopulation where $X = x$ and define the optimal assignment for a given x as:

$$a^*(x) = \underset{j}{\operatorname{argmax}} \mathbb{E}[Y(j)|X = x] \tag{11}$$

Without the argmax, the right side of equation 11 effectively represents the setup of a predictive model. Hence, statistical modeling allows flexibility here by not requiring predefined subpopulations of interest. In Section 2.2.2, I introduce meta-learners, methods capable of estimating a^* from data, with each generating a treatment-assignment policy $\hat{a}(x)$.

One can evaluate treatment assignment policies by assessing how well they minimize the expected difference between the outcomes under optimal assignments, $Y(a^*(X))$, and the outcomes when using policy-driven assignment, $Y(\hat{a}(X))$. In decision theory, this measure is commonly referred to as *(expected) regret*:

$$\text{regret}(\hat{a}) = \mathbb{E}[Y(a^*(X)) - Y(\hat{a}(X))] \tag{12}$$

Minimizing the regret is equivalent to maximizing the expected outcome of implementing the policy.

Evaluating treatment-assignment policies using observational data, as is done when training typical machine-learning models, presents challenges because we observe only one potential outcome per individual, the one for the assigned treatment. Consequently, if a policy recommends a treatment that differs from the one assigned in the historical data, we lack the corresponding potential outcome. Nevertheless, with a dataset of n individuals from a randomized A/B test, it is possible derive an unbiased estimate of Equation 12 (Li et al. 2010):

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\hat{a}(x_i) = t_i)\frac{y_i}{P(T = T_i)} \tag{13}$$

In this expression, for each individual i, X_i represents the feature vector, t_i the treatment assigned in the data, y_i the observed outcome, and $P(T = t_i)$ the probability of assignment to treatment t_i in the data which is a known value when the data is gathered from a randomized A/B test.

Inspired by Manski (2004), there is a vast amount of literature on developing treatment assignment rules in econometrics (e.g., Hirano and Porter 2009; Tetenov 2012; Kitagawa and Tetenov 2018; Manski and Tetenov 2019; Athey and Wager 2020). Manski (2004) addresses the treatment choice problem for discrete covariates by applying statistical decision theory, introducing a Conditional Empirical Success (CES) rule that maximizes the sample-based welfare function. Building on Manski's work, Hirano and Porter (2009) propose a regression-based assignment rule, and prove its asymptotic optimality through the framework of a limit normal experiment. Under constraints on feasible assignment policies, Kitagawa and Tetenov (2018) introduce Empirical Welfare Maximization (EWM), a generalization of the CES approach. This method first estimates the average outcome function from the data, then chooses a policy as a maximizer within the restricted set of policies as the treatment assignment rule. In this method, Kitagawa and Tetenov (2018) concentrate on a rule that maximizes the inverse propensity score weighted (**IPW**) estimate of the average outcome function. In contrast, Athey and Wager (2020) introduce a rule that maximizes the doubly robust (**DR**) estimate.

On the other hand, statistical inference on treatment assignment policies has received less focus than estimation, though some notable studies to address it. Armstrong and Shen (2015) examine inference to identify individuals who should receive treatment under the optimal assignment policy. Using a multiple hypothesis testing approach, they define a random set of characteristics for which the data strongly indicate a positive conditional average treatment effect. Conversely, Luedtke and Van Der Laan (2016) and Andrews et al. (2024) investigate inference for the average outcome, proposing a general conditional inference framework for parameters chosen as data-based maximizers of a specific criterion. They apply this framework to the EWM problem, enabling valid inference of the average outcome that would result if the EWM-selected policy were implemented. Additionally, Andrews et al. (2024) offer a conditional inference method that conditions on the estimated optimal rule.

In contrast with the methods above in this paper I follow a different approach. I attempt to achieve optimal treatment assignment by estimating heterogeneous treatment effects (HTE). To estimate the HTE, I follow the approach described by Hitsch et al. (2024). This approach focuses on cases in which the predicted optimal treatment coincidentally aligns with the actual, randomly assigned treatment (matched observations). If the average outcomes of these matched observations significantly exceed the mean outcome for the treated group, this suggests that the models are effectively assigning optimal treatments. I explain the strategy further in Section 5.3.

Hitsch et al. (2024) discusses Simester et al. (2020) as one of their primary influences for the recent interest in heterogeneous treatment effects, focusing on evaluating

marketing policies. Conventionally, one would compare two or more proposed targeting policies with randomization by policy at the cost of not being able to use the data for comparison with other policies afterwards. Interest in randomization by action makes alternative policy evaluation possible and paves the way for treatment effect estimation.

2.2.2 Meta-Learner Algorithms

As described in Section 2.1.1, the causal inference literature typically emphasizes estimating aggregate causal effects such as ATE, which reflects the average impact of a treatment across a well-defined population. However, the ATE does not assist in assigning different treatments to individuals, as it offers no differentiation among individuals in the population. In Section 2.1.2, I show that this work is fundamentally motivated by the concept of heterogeneous treatment effects (HTEs), which capture the extent to which a treatment's impact varies among individuals.

To account for HTE, one can estimate conditional average treatment effects (CATEs), which represent the average causal effect given a set of observed features. Consequently, as individuals in the population differ by their features (assuming these features correlate with causal effects), we can estimate distinct causal effects for each individual. While treatment effects may still vary among individuals with the same features (due to unobserved factors influencing the causal effect), estimating HTEs through CATEs enables tailored interventions for individuals without requiring predefined subpopulations. In this framework, we want to estimate $\tau(x) = E[Y_i^1 - Y_i^0|X = x] = E[\tau_i|X]$, or, $E[\delta Y_i(t)|X]$ in the continuous case. In other words, we aim to understand how sensitive units are to the treatment. To enable identification of $\tau(x)$, we have to assume unconfoundedness, meaning that treatment assignment is effectively randomized once we control for the features (Rosenbaum and Rubin 1983).

The concepts underlying CATE estimation have been crucial in advancing methods to learn optimal treatment assignment policies from observational data. Therefore, I formally introduce meta-learner algorithms. A meta-learner provides a framework for estimating the CATE by employing various machine learning estimators, known as base learners (i.e., regression estimators). A meta-learner can employ a single base learner that includes the treatment indicator as a feature (as in the S-learner) or use separate base learners for the treatment and control groups (as in the T-learner, X-learner, and R-learner).

In this paper, I review the two main and most popular classes of meta-learners: X- and R-learners. Several additional meta-learners exist in the literature; however, they fall outside the scope of this paper. This categorization is important because it highlights that outcome prediction, causal effect estimation, and treatment assign-

ment represent distinct tasks that carry significant implications for the application of predictive models in policy-making.

X-Learner

Künzel et al. (2019) introduce the X-learner, a meta-learner that demonstrates high efficiency in estimating CATE when the number of units in one treatment group is significantly larger than that in the other. This learner also leverages structural properties of the CATE function. For instance, if the CATE function is linear and the response functions in both treatment and control groups are (Lipschitz) continuous, the X-learner can still reach the parametric rate under certain regularity conditions. In addition, the authors introduce various X-learner variants that employ random forests (RF) and BART model as base learners.

The core concept of the X-learner unfolds in three stages. First, it estimates the response functions:

$$\mu_0 = \mathbb{E}[Y(0)|X = x]$$
$$\mu_1 = \mathbb{E}[Y(1)|X = x]$$

using any supervised learning or regression algorithm, with the resulting estimates denoted as $\hat{\mu}_0$ and $\hat{\mu}_1$. These algorithms are called the base learners for the first stage.

Second, the method imputes the treatment effects for individuals in the treated group using the control-outcome estimator, and for individuals in the control group using the treatment-outcome estimator. That is expressed as:

$$\tilde{D}_i^1 := Y_i^1 - \hat{\mu}_0(X_i^1)$$
$$\tilde{D}_i^0 := \hat{\mu}_1(X_i^0) - Y_i^0$$

The authors call these values the imputed treatment effects. In this case, if $\hat{\mu}_0 = \mu_0$ and $\hat{\mu}_1 = \mu_1$, then $\tau(x) = \mathbb{E}[\tilde{D}^1|X = x] = \mathbb{E}[\tilde{D}^0|X = x]$. To estimate $\tau(x)$ can be used any supervised learning or regression, in two ways: by applying the imputed treatment effects as the response variable in the treatment group to obtain $\hat{\tau}_1(x)$, and similarly in the control group to obtain $\hat{\tau}_0(x)$. These algorithms are referred to as the second stage base learners.

The third stage includes defining the CATE estimate, a weighted average of two estimates obtained in stage two:

$$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x) \tag{14}$$

where g is a weight function, $g \in [0, 1]$

Künzel et al. (2019) remark that both $\hat{\tau}_0$ and $\hat{\tau}_1$ serve as estimators for τ, with g selected to combine them into a refined estimator, $\hat{\tau}$. Based on their experience, using an estimate of the propensity score for g, such as $g = \hat{e}$, often works well. However, it may also be appropriate to set $g = 1$ or $g = 0$ if the treated unit count is substantially larger or smaller than that of the control units. For certain estimators, estimating the covariance matrix of $\hat{\tau}_1$ and $\hat{\tau}_0$ may be possible, and allow g to be selected to minimize the variance of $\hat{\tau}$.

Although the X-learner performs effectively with randomized treatments, it struggles to construct counterfactuals from X_i in observational studies (Künzel et al. 2019). Causal forests address this limitation by using $g(X)$ as the propensity score, enabling CATE estimation through a doubly robust (DR) approach. Due to these limitations, in the application section of this paper, I focus on using the R-learner instead.

R-Learner

Nie and Wager (2021) propose the R-learner, a new approach to estimating heterogeneous treatment effects that offers a comprehensive answer to how machine learning methods should be adapted for treatment effect estimation in observational studies. First, the authors express treatment propensity as (similar as in the Equation 4):

$$e^*(x) = P(T = 1|X = x) \tag{15}$$

where '$*$' superscript denotes unknown population quantities. Here, I use the same notation as in the previous sections of this paper. Then, assuming unconfoundedness:

$$E\{\varepsilon_i(T_i)|X_i, T_i\} = 0 \tag{16}$$

$$\varepsilon_i(T) := Y_i(t) - \{\mu^*_{(0)}(X_i) + t\tau^*(X_i)\} \tag{17}$$

To express the CATE function $\tau^*(x)$ in terms of the *conditional mean outcome,* $m^*(x)$, first I express $m^*(x)$ as:

$$m^*(x) = E[Y|X = x] = \tau^*_{(0)}(X_i) + e^*(X_i)\tau^*(X_i) \tag{18}$$

Then, using Robinson (1988) expressions, the CATE function can be written as:

$$Y_i - m^*(X_i) = \{T_i - e^*(X_i)\}\tau^*(X_i) + \varepsilon_i \tag{19}$$

, where $\varepsilon_i = \varepsilon_i(T_i)$.

Robinson (1988) first applied this decomposition to estimate parametric components within partially linear models. Athey et al. (2019) later leveraged it to develop a causal forest that addresses confounding, while Chernozhukov et al. (2018) used it as a prime example of how machine learning methods effectively estimate nuisance components in semiparametric inference. However, these results primarily focus on estimating parametric models for $\tau(\cdot)$ or, in the case of Athey et al. (2019), on local parametric modelling.

Nie and Wager (2021) demonstrate how the Robinson transformation (Equation 19) enables flexible treatment effect estimation using modern machine learning techniques, including boosting (e.g., xgboost, random forests, and causal forests) and deep learning (e.g., feed-forward neural networks). They show that this approach allows construction of a loss function that captures heterogeneous treatment effects, and that treatment effects can then be estimated accurately, both in empirical performance and with asymptotic guarantees, by identifying regularized minimizers of this loss function.

Equation 19 can be rewritten as (Robins 2004):

$$\tau^*(\cdot) = \underset{\tau}{\operatorname{argmin}}\{E([\{Y_i - m*(X_i)\} - \{T_i - e^*(X_i)\}]^2)\} \tag{20}$$

In this case, an oracle with prior knowledge about both functions $m(x)$ and $e(x)$ could estimate HTE function $\tau(x)$ through loss function minimization $\tilde{\tau}(\cdot)$ (Nie and Wager 2021):

$$\tilde{\tau}(\cdot) = \left(\frac{1}{n} \sum_{i=1}^{n} [\{Y_i - m*(X_i)\} - \{T_i - e^*(X_i)\}\tau(X_i)]^2 + \Lambda_n\{\tau(\cdot)\} \right) \tag{21}$$

where $\Lambda\{\tau(\cdot)\}$ acts as a regularizer on the complexity of the $\tau(x)$ function. This regularization might be explicit, as seen in penalized regression, or implicit, as in a well-constructed deep neural network. However, in practice, we rarely know the weighted main effect function $m(x)$ and often lack knowledge of treatment propensities $e(x)$, making the estimator in Equation 21 infeasible (Nie and Wager 2021).

Based on these preliminaries, Nie and Wager (2021) develop a class of two-step estimators using **cross-validation** (or cross-fitting). In the first step, they fit $\hat{m}(x)$ and $\hat{e}(x)$ with machine learning methods using cross-validation by dividing up the data into Q equal sized folds. Then, in the second step, they estimate treatment effects by minimizing the **R-loss** function, $\hat{L}_n(\tau(x))$:

$$\hat{L}_n(\tau(x)) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(Y_i - \hat{m}^{(-i)}(X_i)\right) - \left(T_i - \hat{e}^{(-i)}(X_i)\right)\tau(X_i)\right\}^2 \tag{22}$$

where $\hat{e}^{(-i)}(X_i)$ and $\hat{m}^{(-i)}$ denote the out-of-the fold held-out predictions obtained without using i-th training sample. This approach is the **R-learner**. In essence, the first step of the approach approximates the oracle objective, while the second step focuses on optimizing it.

I chose to implement R-learner for the application part of this paper for two main reasons. First, this method eliminates correlations through the structure of the loss function L. Second, it allows the representation of $\tau(x)$ to be shaped by the selected optimization method for loss-function (Equation 22). Because optimizing Equation 22 is an empirical minimization problem, I tackle it by using the tree-based method of Causal Forests (CF) and the feed-forward deep learning method of Causal Neural Networks (CNN). Additionally, through the R-learner technique I can fine-tune these two methods by cross-validating on the loss function L, eliminating the need for more complex model-assisted cross-fitting techniques.

Comparing the meta-learners to highlight a few key points. First, none of the meta-learners is the single best technique, and each has weaknesses. The application of each method is highly dependent on the context of the problem at hand. Second, the choice of base learner can significantly impact the prediction accuracy of the meta-learner. Meta-learners utilize various predictive ML models, including linear regression, boosted decision trees, neural networks, and Gaussian processes. Therefore, the effectiveness of a meta-learner often depends significantly on the choice of machine learning methods it incorporates. Usually, finding the best-performing model requires experimenting with multiple options to determine what works optimally. This flexibility is a valuable advantage for meta-learners, allowing practitioners to leverage domain knowledge when selecting high-performing base learners.

A crucial final point to consider is that optimizing models for causal effects prediction differs from optimizing models for predicting optimal treatment assignments (Fernández-Loría et al. 2023). Figure 2 depicts the contrast of outcome predictions from two models for a single individual. One model produces high prediction errors (Figure 2 (a)), while the other yields low errors (Figure 2 (b)). In both figures, triangles represent true conditional expectations, and dots indicate model predictions. A greater distance between triangles and dots (shown by dashed lines) signals poorer outcome predictions.

In this scenario, the conditional expectation when $T = 1$, is higher than $T = 2$, indicated by triangles, suggesting that $T = 1$ is the better treatment strategy. Thus, models achieve the optimal assignment when $\hat{\mu}(x, 1) > \hat{\mu}(x, 2)$. Figure 2 (a) shows

that, despite its larger prediction errors, the first model correctly makes the optimal treatment assignment because the rank ordering predicted outcomes aligns with the true values ranking. Conversely, the second model in Figure 2 (b) results in a poorer assignment, despite smaller prediction errors, as the ordering is reversed. Notably, this misalignment can also arise when fitting models for causal effect prediction. Therefore, because higher accuracy causal effect prediction may actually worsen treatment assignments, training models focused on predicting optimal assignments should yield better treatment assignments than the other meta-learners.

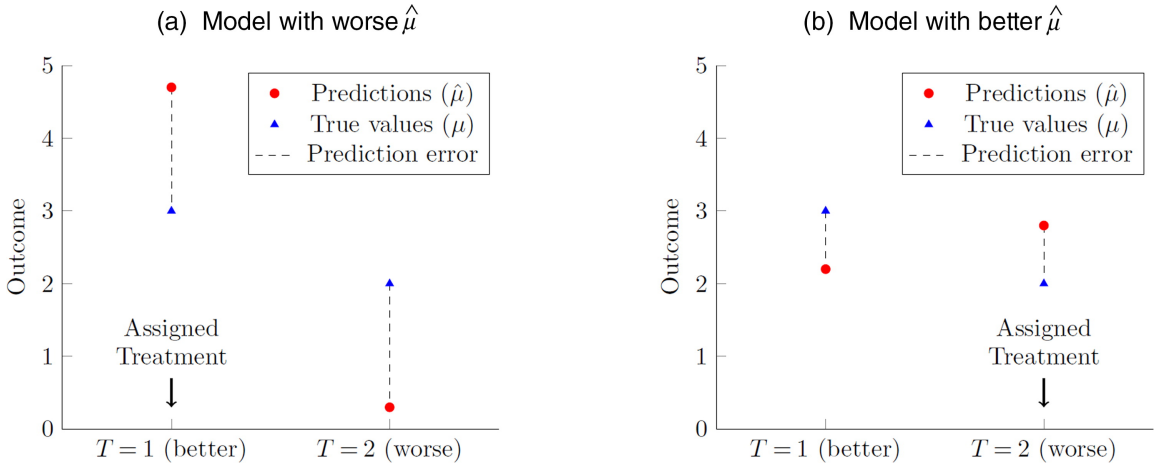

Figure 2: **Source**: Fernández-Loría et al. (2023)
Comparing causal effects and treatment assignment prediction for an individual. The model shown on the left has greater outcome prediction errors compared to the model on right, as indicated by the larger dashed lines on the left than on the right. Nonetheless, the model on the left achieves a better treatment assignment than the model on the right because the dots maintain the ranking order of the triangles.

Machine Learning for Causal Inference

In this section, I will outline the methodology behind heterogeneity and conditional average treatment effect (CATE) estimation using machine learning methods. Hitsch et al. (2024) categorize the estimation methods into two groups, distinguishing them by their conceptual estimation approaches. The first group comprises *indirect estimation methods* that aim to minimize the squared-error loss between observed and predicted outcome levels, $\mathbb{E}[(Y_i - \hat{\mu}(X_i, T_i))^2]$. The best predictor in this approach is the regression function, $\mu(x,t) = \mathbb{E}[Y|X = x, T = t]$.

Given unconfoundedness, we have:

$$\mu(x,t) = \mathbb{E}[Y(t)|X = x, T = t] \tag{23}$$

Therefore, once we have estimated the regression function, we can indirectly predict the CATE, as do (Hitsch et al. 2024):

$$\hat{\tau}(x) = \hat{\tau}(x,1) - \hat{\tau}(x,0) \tag{24}$$

However, indirect estimation may be inefficient, because our primary objective is to predict CATEs rather than outcome levels. Thus, we should ideally use *direct estimation methods* that focus on minimizing loss $\mathbb{E}[(\tau(X_i) - \hat{\tau}(X_i))^2]$. Hitsch et al. (2024) state that this loss function is infeasible due to the fundamental problem of causal inference, making direct estimation of treatment effects appear impossible. Nevertheless, there exist recent methods that attempt to directly estimate the CATE, such as the causal forest approach by Wager and Athey (2018).

In this paper, I use the semiparametric *indirect estimation method* of causal neural networks (CNN) (Farrell et al. 2021), and the parametric *direct estimation method* of causal forests (CF) (Wager and Athey 2018). In the next section, I introduce first ensemble methods such as boosting, focusing on the CF technique. Then, I explain the deep-forward neural network technique of CNN.

3.1 Introduction to Tree-Based Methods

Wager and Athey (2018) extend random forest framework of Breiman (2001) to estimate treatment effects directly, eliminating the need to construct counterfactual outcomes beforehand. In traditional random forests, the algorithm repeatedly splits the data to minimize the prediction error of the outcome variable. Causal forests operate similarly but differ in their splitting criterion: instead of minimizing prediction error, they parti-

tion the data to maximize differences between the outcome variable and the treatment variable across the splits.

3.1.1 Causal Trees

A causal forest, much like a random forest, is composed of multiple decision trees, in this case, causal trees. Introduced by Athey and Imbens (2016), causal trees are adaptations of regression trees designed for causal inference. Regression trees are decision trees used to predict a continuous outcome variable Y_i by creating a piecewise constant approximation of the data. They recursively partition the data by making binary splits based on one variable X_j at a time, resulting in rectangular regions R_t for $t = 1, \ldots, T$, where T is the total number of regions. Figure 3 illustrates an example of such a regression tree and its corresponding regions.

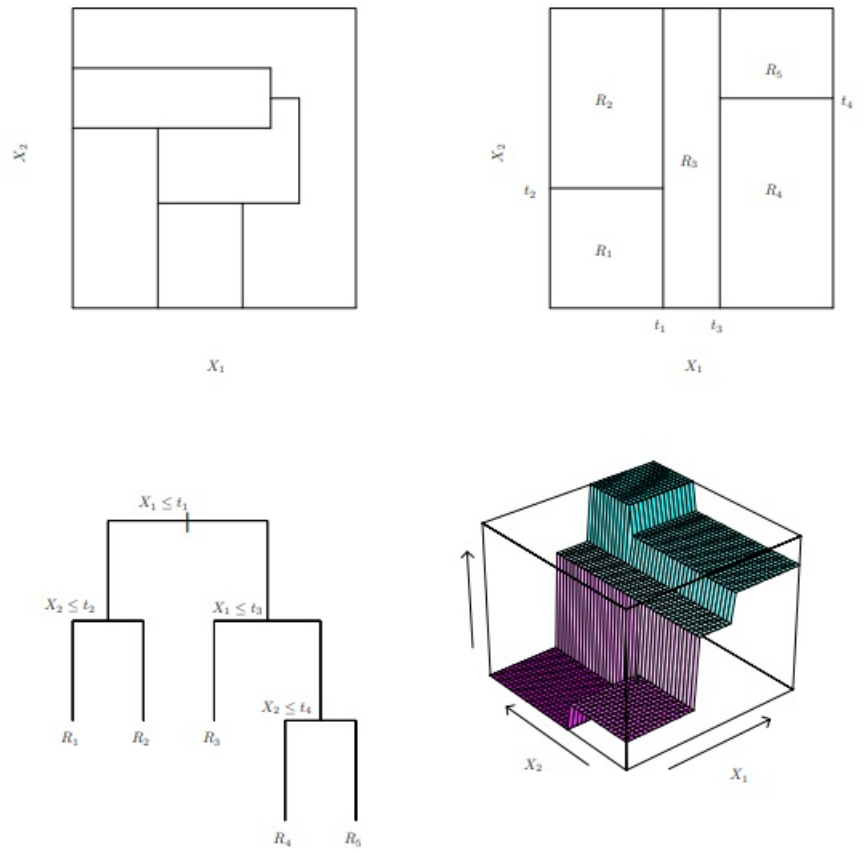

Figure 3: **Source**: (James et al. 2013, p. 335)
Top Left: A partition of the two-dimensional feature space (i.e., using two covariates) that cannot be generated by recursive binary splitting. Top Right: The result of applying recursive binary splitting to a two-dimensional example. Bottom Left: A decision tree representing the partition displayed in the top right panel. Bottom Right: A perspective plot of the prediction surface associated with that decision tree.

In causal trees, the primary objective is similar to that of regression trees: minimize the prediction error, specifically $\sum_{i=1}^{N}(\tau_i - \tau(x))^2$, where τ_i is the true treatment effect and $\tau(x)$ is the estimated treatment effect based on features X_i. However, because the true treatment effects τ_i are unobserved, constructing the tree structure and deciding where to split the nodes is not straightforward. To address this challenge, two main approaches are employed: outcome transformation and objective function transformation. The former transforms the outcome variable Y_i and then applies standard regression tree methods directly. The latter modifies the objective function used for splitting the tree, leading to what is called *transformed objective trees*.

A significant contribution of Athey and Imbens (2016) is the development of methods to construct confidence intervals and conduct hypothesis tests within this framework. They introduced the concept of *honesty*, which ensures consistency and asymptotic normality of the treatment effect estimates without requiring additional assumptions. The idea behind honesty is to split the training data into two separate subsets: S^{tr} used solely to determine the tree structure by splitting the data into nodes; S^{est} used solely to estimate the treatment effects within each node. The test data is denoted by S^{te}. By separating the data used for splitting from the data used for estimation, the asymptotic properties of the treatment effect estimates within each leaf are preserved, as if the leaves were endogenously given rather than influenced by the data.

Because the true individual treatment effects τ_i are unobservable, we cannot directly compute an error function like the mean squared error. To overcome this, transformed objective trees use alternative objectives that focus on maximizing the heterogeneity of treatment effects between different leaves. These objectives may also favor splits that reduce variance within leaves, thereby improving the reliability of the estimates.

Athey and Imbens (2016) derive a primary splitting criterion based on maximizing the negative expected value of a modified mean squared error (EMSE) for a given tree partition Π [4]:

$$
\begin{aligned}
-\widehat{EMSE}_{\tau}(S^{tr}, N^{est}, \Pi) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \hat{\tau}^2(X_i; S^{tr}, \Pi) \\
- \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \cdot \sum_{R_t \in \Pi} \left(\frac{S^2_{S^{tr}_{treat}}(R_t)}{p} + \frac{S^2_{S^{tr}_{control}}(R_t)}{1-p} \right) \quad (25)
\end{aligned}
$$

In this equation, $\hat{\tau}^2(X_i; S^{tr}, \Pi)$ is the estimated treatment effect for X_i in the tree building sample S^{tr}, given the candidate tree structure Π. N^{tr} and N^{est} are the numbers of observations in the tree-building S^{tr} and treatment effect estimating S^{est} samples,

[4] Π stands for the full tree structure, including all splits.

respectively. $S^2_{S^{tr}_{treat}}$ and $S^2_{S^{tr}_{control}}$ are the within-leaf variances for treated and control observations in leaf R_t, calculated from tree building sample S^{tr} only. Lastly, p represents the proportion of treated observations in the entire training sample.

The first term in this criterion rewards splits that increase the heterogeneity of treatment effects across different leaves, basically increasing the differences in estimated effects between nodes. The second term penalizes splits that result in high variance within leaves, encouraging more precise estimates within each node. By balancing these two aspects, maximizing between-leaf heterogeneity and minimizing within-leaf variance, the splitting criterion aims to create a tree structure that captures variations in treatment effects while maintaining reliable estimates within each group.

In their later work on generalized random forests, Athey et al. (2019) propose additional splitting criteria to further improve the method's flexibility to a wider range of problems.

3.1.2 From Causal Trees to Forests

The process of extending causal trees to causal forests mirrors that of regression trees to random forests. To build a causal forest, a predetermined number of bootstrap samples are drawn with replacement from the training dataset. Then, an individual causal tree is trained on each of these bootstrap samples.

To reduce the correlation among the trees and improve the model's performance, each tree considers only a random subset of covariates (features) at each split decision. This subset is smaller than or equal to the total number of covariates. This random selection of features at each split helps to create diverse trees that, when combined, improve overall prediction accuracy.

The final prediction of the causal forest is obtained by averaging the predicted treatment effects from all the individual causal trees. This ensemble prediction is calculated using the formula:

$$\hat{\tau}_{forest}(X) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}^b(X) \tag{26}$$

,where $\hat{\tau}^b(X)$ is the predicted treatment effect from the b-th causal tree trained on the b-th bootstrap sample. By averaging the predictions from multiple diverse trees, the causal forest provides a more accurate estimate of the treatment effect compared to a single causal tree.

3.2 Background of the Deep Learning Model

Farrell et al. (2021) introduce a semiparametric deep learning method called causal

neural networks (CNN) for estimating heterogeneous treatment effects (HTE). This method uses a feed-forward neural network to jointly estimate both the treatment effect and the impact of covariates i.e., approximating the functions $\alpha(x)$ and $\beta(x)$:

$$\mathbb{E}[Y_i|X_i = x, T_i = t] = \alpha(x) + \beta(x) \times t$$

Historically, neural networks have been less commonly employed in economic research than other machine learning techniques. In the following sections, I will introduce the fundamentals of feed-forward neural networks and describe how the network architecture is adjusted to construct a causal model for predicting HTE.

3.2.1 Basics of Feed-Forward Neural Networks

A feed-forward neural network is a type of non-linear model used to approximate a target function $f(x)$. The term *feed forward* indicates that the connections between the nodes form a directed acyclic graph, where information moves in one direction from the input layer, through any hidden layers, to the output layer, without forming any loops. The input layer consists of J nodes, where J is the number of input variables or covariates. Each hidden layer contains a predetermined number of nodes, known as the *width* of that layer.

The computation within the network proceeds as follows: First, for a node u in hidden layer l, the input $z_u^{(l)}$ is calculated by summing the weighted outputs of all nodes in the previous layer:

$$z_u^{(l)} \equiv \sum_{b=1}^{(l-1)} \beta_{bu}^{(l-1)} a_b \tag{27}$$

,where $\beta_{bu}^{(l-1)}$ is the weight connecting node b in layer $(l-1)$ to node u in layer l. Moreover, $a_b^{(l-1)}$ is the output of node b in the previous layer $(l-1)$. Then, ReLU [5] activation function is applied to the input $z_u^{(l)}$ to introduce non-linearity for the output node $a_u^{(l)}$:

$$a_u^{(l)} = \max(z_u^{(l)}, 0)$$

In other words, the output $a_u^{(l)}$ is zero if $z_u^{(l)} < 0$, and equal to $z_u^{(l)}$ if $z_u^{(l)} > 0$.

Moreover, no activation function is applied to the input and output layers so:

[5]The rectified linear unit (ReLU) activation function introduces the property of nonlinearity to a deep learning model and solves the vanishing gradients issue.

$$a_b^{(l=1)} = x_b$$
$$a^{(l=L)} = \hat{y}$$

Furthermore, neural networks can produce prediction of multiple outcomes simultaneously, so $a_b^{l=L} = \hat{Y}_b$, however, in this paper I focus on predicting only one outcome variable at the time.

Farrell et al. (2021) emphasize the use of only ReLU activation function $x \mapsto max(x, 0)$ in their approach. They argue that the shift from traditional smooth, sigmoid type activation functions to the ReLU function is a key reason, alongside advancements in computational power and optimization techniques, for the recent surge in the performance of neural networks. The ReLU function addresses issues like vanishing gradients and has been shown to outperform earlier activation functions like sigmoid or tanh.

Furthermore, the network's weights $\beta_{bu}^{(l)}$ are optimized using the *back-propagation* technique, which iteratively updates the weights to minimize prediction error. The main idea is that, for each prediction, the algorithm computes the error based on a chosen loss function (e.g., mean squared error (MSE)). Then, it calculates how much each weight contributes to this error by computing the gradient of the loss function with respect to each weight. Finally, the weights are adjusted in the opposite direction of the gradient (gradient descent), and both their magnitude and sign are modified to reduce the overall error. By repeatedly performing these steps over multiple iterations (*epochs*), the network learns the optimal weights that best approximate the target function $f(x)$, leading to improved predictive performance.

3.2.2 Architecture of Causal Deep Neural Network

The causal neural network (CNN) introduced by Farrell et al. (2021) is a feed-forward neural network that employs ReLU activation functions and incorporates, an input layer, two custom layers appended to the hidden layers, a parameter layer and an output layer. A visual representation of this network architecture with two hidden layers is illustrated in Figure 4.

In this CNN architecture, the second-to-last layer, referred to as the parameter layer (L_4 in Figure 4), consists of three nodes. Two of these nodes are standard nodes that receive their inputs from the last hidden layer through feed-forwarding, with no activation function applied. The third node is the treatment indicator T_i, which is not connected to the previous hidden layers and effectively acts as an additional input node.

The final output layer is designed to predict the outcome variable Y_i. This prediction is computed by summing the outputs of the two standard nodes from the parameter

layer, with one of these node outputs being multiplied by the binary treatment indicator T_i. Specifically, the node that interacts with T_i estimates the heterogeneous treatment effect $\hat{\tau}(X)$, while the node that does not interact with T_i estimates the outcome under no treatment $\hat{Y}^0$.

The network is trained to minimize the mean squared error (MSE) between the predicted outcomes and the actual outcomes. This involves jointly estimating $Y^0(X)$ and $\tau(X)$ (and consequently $Y^1(X)$) by solving the following optimization problem:

$$\begin{pmatrix} \hat{Y}^0(X) \\ \hat{\tau}(X) = \hat{Y}^1(X) - \hat{Y}^0(X) \end{pmatrix} := \underset{\tilde{Y}^0, \hat{\tau}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{Y}^0(X_i) - \tilde{\tau}(X_i)T_i)^2 \tag{28}$$

Here, y_i represents the observed outcome for individual i, and n is the total number of observations. Note that MSE serves as the loss function that quantifies the average squared difference between the predicted values and the actual outcomes, therefore measuring the magnitude of the prediction error.

To generate predictions for the CATE or HTE, the model feed-forwards the covariates through the network and extracts the value of $\hat{\tau}(X)$ from the parameter layer, the node that interacts with treatment indicator T_i. According to Farrell et al. (2021), this approach of jointly estimating $Y^0(X)$ and $\tau(X)$ within the same neural network outperforms methods that separately estimate $Y^0(X)$ and $Y^1(X)$ using neural networks trained on different subsets of the data.

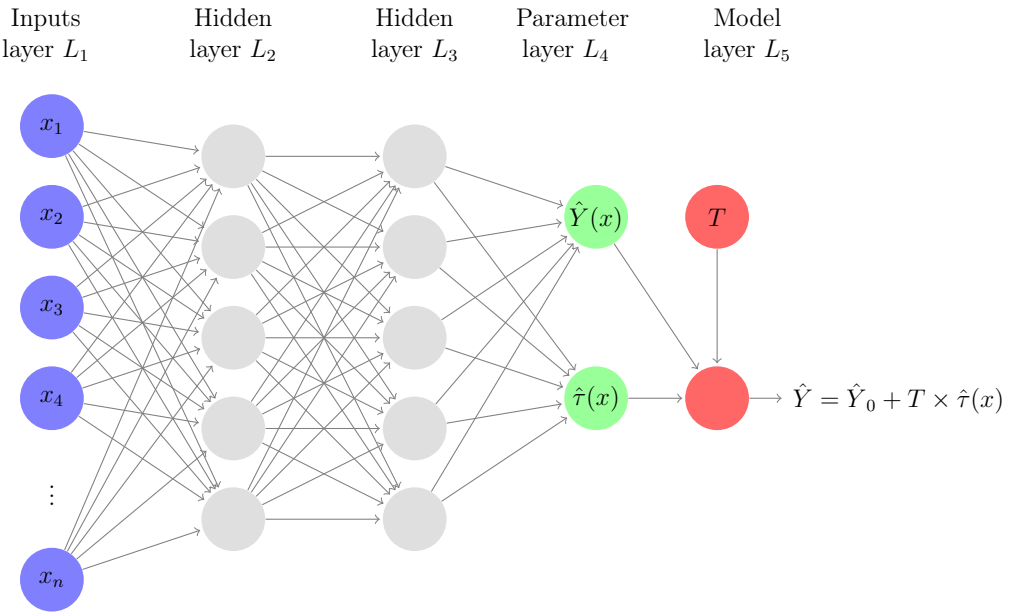

Figure 4: **Source**: Farrell et al. (2021). This figure shows the feed-forward neural network architecture with two hidden layers. I employ the same structure when constructing the causal neural network (CNN) in the application section of this paper.

Empirical Application

4.1 Background

In recent years, estimation of heterogeneous causal effects has garnered significant attention across various research disciplines. Machine learning (ML) methods have been adapted to flexibly estimate heterogeneity (HTE) along potentially large numbers of covariates, leading to the development of estimators utilizing random forests (Wager and Athey 2018; Athey et al. 2019), LASSO (Tian et al. 2014; Chen et al. 2017), deep neural networks (Johansson et al. 2016; Schwab et al. 2018), and Bayesian ML approaches (Taddy et al. 2016).

Despite advancements in estimating average causal effects, practical guidance for practitioners on estimating HTEs remains limited, particularly in the context of optimal treatment assignment. The economic literature on causal inference for optimal treatment assignment has gained popularity recently, especially in the medical sector, where there is growing interest in patient-centered outcomes (Willke et al. 2012). However, significant potential exists for applications in other fields, such as incentivizing effort in employee management (DellaVigna and Pope 2018) and evaluating the effectiveness of public policies in education and taxation (Johansen 2024; Xu et al. 2024). Kleinberg et al. (2015) clarify the distinction between the need for causal inference and prediction in policy applications. Kitagawa and Tetenov (2018) develop a frequentist empirical welfare maximization method for optimal treatment assignment, while Manski (2004) suggests that optimal treatment assignment to maximize social welfare differs from traditional point estimation with hypothesis testing. Building on this, Hirano and Porter (2009) develop an asymptotic normality theory for statistical treatment rules that map empirical data into treatment choices.

Classical nonparametric approaches for estimating HTEs, such as kernel methods and nearest-neighbor matching, provide good predictive performance with few covariates, but experience a rapid decline in accuracy as the number of covariates increases (Wager and Athey 2018). This limitation underscores the argument for employing ML methods, which tend to perform better with many covariates, but often require a larger number of observations to produce reliable predictions.

Given the rapidly expanding literature on ML methods for HTE estimation, an important question arises on which methods perform well for optimal treatment assignment, and how can they be compared empirically.

A central goal of this paper is to compare the performance of two machine learning models in predicting heterogeneous treatment effects for optimal treatment assignment, representing the two main categories of estimation methods: direct and indirect.

Specifically, I compare the indirect semiparametric estimation method of causal neural networks (CNN) by Farrell et al. (2021) with the direct parametric estimation method of tree-based causal forests (CF) from Wager and Athey (2018).

I apply these ML methods to an empirical dataset from an online experiment on the incentivization of manual labor conducted by Opitz et al. (2024), using a comparison method based on Hitsch et al. (2024), and the loss function **R-learner** developed from Nie and Wager (2021). The data was collected from an experiment carried out on Amazon Mechanical Turk (MTurk), a platform used primarily for small-scale contract labor but increasingly popular for behavioral experiments.

A major challenge in optimal treatment assignment using HTEs is the *winner's curse*, in which overestimated treatments are more likely to be identified as optimal because the selection is based on the highest predicted treatment effect. To address this issue, I present shrinkage estimators as a possible solution, and evaluate the effectiveness of these shrinkage techniques when applied to the predictions of the ML methods on the empirical dataset.

This study is closely related to the work of Opitz et al. (2024) and Hitsch et al. (2024). Opitz et al. (2024) examines the performance of targeted assignment of incentive schemes by conducting two large-scale experiments, each involving an extensive personality trait survey followed by a manual labor task. The first experiment was used for pre-analysis, model selection, and training of the model, while the second compared the performance of treatment assignment using the Virtual Twin Random Forest method. Opitz et al. (2024) found that personality traits could predict participants' performance, enabling employers to exploit worker heterogeneity to enhance the performance effect of incentives through targeted assignment.

However, this study differs from Opitz et al. (2024) and the rest of the related literature in two major aspects. First, I aim to empirically compare two ML methods for optimal treatment assignment, and to offer a rigorous guideline on how to compare and select methods for optimal treatment assignment. Second, I analyze the critical issue of the winner's curse in optimal targeting, which was not highlighted in Opitz et al. (2024), and not widely developed in the literature of the optimal targeting. I introduce two shrinkage estimators as a solution, offering a novel setting for the application of such techniques for improved optimal treatment assignment.

Although I employ a methodology similar to that used by Hitsch et al. (2024) to compare ML methods, this study differs in key respects. Hitsch et al. (2024) focuses on customer targeting in marketing contexts, such as companies using catalogs, emails, and display ads, while this empirical application centers on incentive schemes in the context of workers. Additionally, Hitsch et al. (2024) emphasizes two main techniques,

that of treatment effect projection (TEP) and causal KNN regression, whereas I focus on CF and CNN methods.

The contributions of this paper to the literature are twofold. First, I rigorously compare the performance of two machine learning methods in estimating heterogeneous treatment effects for optimal policy assignment, and provide insights into their relative effectiveness. Second, I introduce the use of shrinkage estimators in a novel setting, in the context of estimating HTE for optimal treatment assignment, addressing the winner's curse problem.

4.2 Data

In this paper, I analyze data [6] from the first round of experiments conducted by Opitz et al. (2024) on Amazon Mechanical Turk (MTurk) with a sample exclusively from the United States, collected over approximately two and a half weeks in September 2021. This paper focuses on using the data collected in the initial round of the experiment to apply and compare the machine learning models for optimal treatment assignment. The dataset consists of 6,065 observations (individuals) and 55 features, where each individual is assigned to one of the six treatments or control group.

Before participating in the main task of the experiment, participants completed an extensive survey covering their demographics, personality traits, and social preferences. In addition to standard questions about age, gender, and education level, the survey included assessments of the Big Five personality traits (John 1991), as well as measures of risk preferences, loss aversion, competitiveness, social comparison, altruism, and positive reciprocity.

For the working task, participants could earn points by alternately pressing 'a' and 'b' buttons over a ten-minute period. *The participants are asked to try to score as many points as possible.* Then, the points participants score in this test serve as **a proxy for ability** in such tasks. During this task, they could see a timer, their current point total, and their current bonus. Note that the essential aspect of an experiment is to achieve randomization by employing a randomization strategy. Each participant was randomly assigned to one of six treatments or to a control group. The assignment determined the incentive structure, or lack thereof, for pressing the buttons and accumulating points.

These treatments were:

1. **Pay for Performance (PfP)** 5 cents for every 100 points.

2. **Goal** $1 if you score at least 2000 points.

[6] The data files are retrieved from the supplemental material webpage of Opitz et al. (2024): https://doi.org/10.1287/mnsc.2022.03362

3. **Gift & Goal** \$1, would appreciate at least 2000 points appreciate if you try to score at least 2,000 points.

4. **Loss** \$1, lose it unless you score at least 2000 points.

5. **Real-Time Feedback** \$0.02 times the percentile reached.

6. **Social PfP** 3 cents for participant + 2 cents for Doctors without Borders for every 100 points.

7. **Control** no extra payment.

For detailed descriptions of the treatments and the text shown to the participants, see Appendix 9.2 or Opitz et al. (2024).

In all treatment groups, participants earned significantly more points than those in the control group (see Table 1: Wilcoxon Rank Sum test with $\alpha < 1\%$). The mean and median outcomes for each group are illustrated in Figure 5.

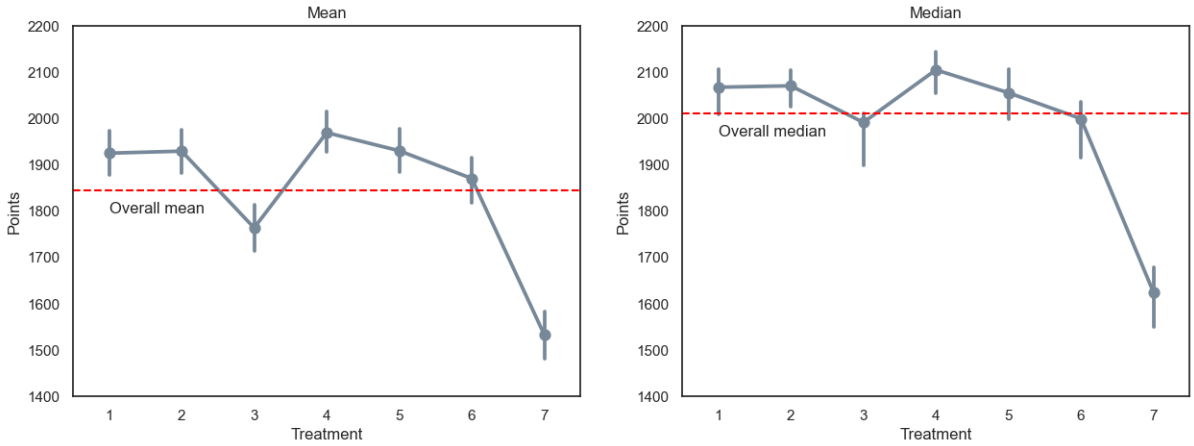

Figure 5: These graphs depict the mean and median outcomes for the respective treatments. The enumeration of treatments is the same as the order listed in Appendix 9.2, with treatment seven being the control group. The overall mean of points is 1845, and the overall median is 2,012. The bars represent the bootstrap confidence intervals at the 95% level

Importantly, the average outcomes for treatments one, two, four, and five were significantly higher than those for treatments three and six (based on Wilcoxon Rank Sum tests at the 5% significance level; see Table 1). This suggests that certain treatments were more effective in motivating participants.

Figure 6 illustrates the estimated probability density functions of the points earned in each treatment. Notably, there is a spike in density around zero points, especially in the control group and treatment three. In these groups, the payoff did not depend on

	T1	T2	T3	T4	T5	T6	T7
T1	0.500	0.554	0	0.90	0.534	0.046	0
T2	0.445	0.500	0	0.894	0.502	0.026	0
T3	0.99	0.999	0.5	1.000	0.999	0.993	0
T4	0.098	0.105	0	0.500	0.135	0.0009	0
T5	0.465	0.497	0	0.864	0.500	0.047	0
T6	0.953	0.973	0.006	0.999	0.952	0.5	0
T7	1	1	1	1	1	1	0.5

Table 1: This table contains the p-values for the Wilcoxon Rank Sum tests for the hypothesis $H_0: P(X > Y) = P(Y > X)$ *vs* $H_1: P(X > Y) > P(Y > X)$ with X being the points of the sample where the treatment corresponds to the treatment of the respective row value, and Y being the points of the sample where the treatment corresponds to the treatment of the column value. Treatment seven is the control group. p-values that are less than 0.05 are colored in red.

the number of points scored, which explains the high frequency of zero-point outcomes and the substantial gap between the mean and median.

To ensure data quality, the dataset was cleaned to remove participants who did not engage correctly with the experiment. This exclusion was based on predefined criteria, such as not pressing any buttons, spending less than a certain amount of time on the task pages, or achieving point totals indicative of cheating.

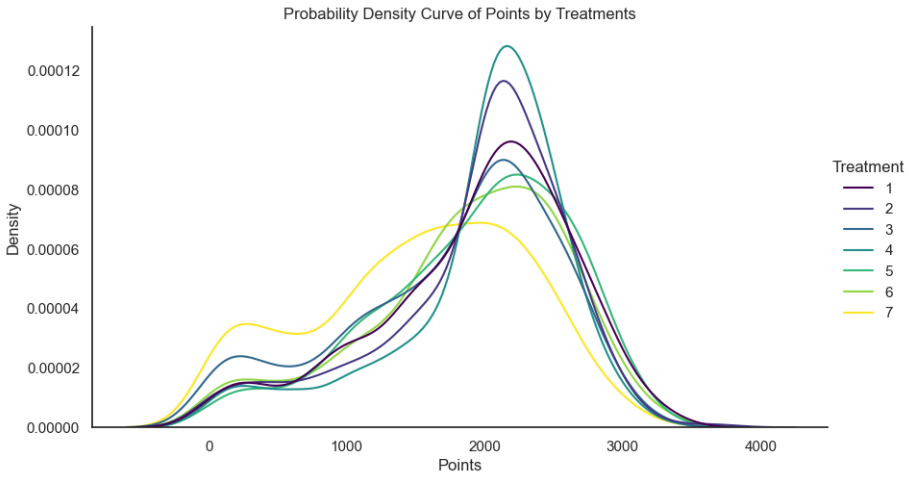

Figure 6: This figure depicts the estimated kernel density functions for the points for each treatment. The functions are colored according to the six different treatments to be taken from the adjacent legend. The enumeration of treatments is the same as the order listed in Appendix 9.2.

Figure 7 illustrates the average score accumulated for each treatment. The figure suggests that "Loss" (or treatment 4) is the most effective treatment, followed by the

"Real-Time Feedback", "Goal" and "Pay for Performance". In contrast, the control group shows intuitively the lowest accumulated of scored point, which translates to individuals having the lowest performance.

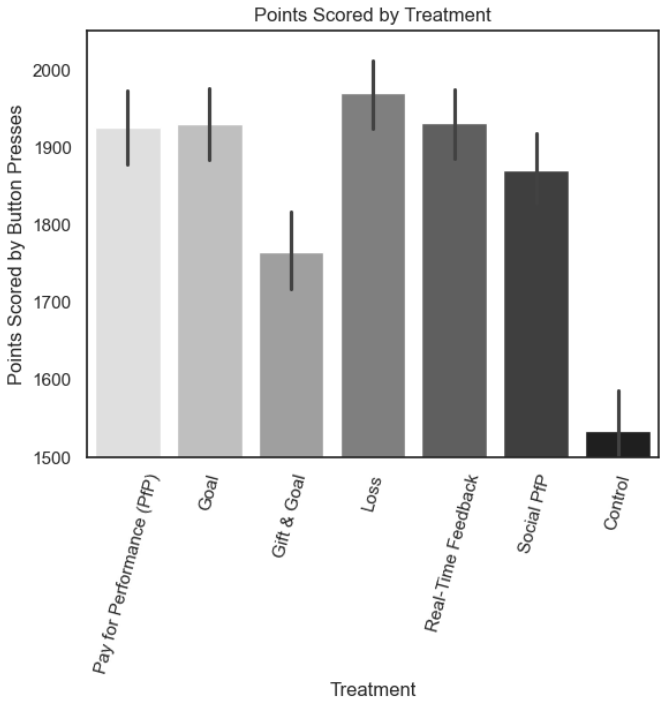

Figure 7: The figure displays the average individual performance categorized by treatment group, with treatment details provided in Appendix 9.2. Performance is quantified by points scored in the button pressing task. Vertical lines indicate the 95% confidence interval.

Methodology

5.1 Model Training and Hyperparameter Tuning

To determine the optimal hyperparameters for both the causal forest and causal neural network model, I employ a grid search algorithm combined with three-fold cross-validation [7]. Specific details about which parameters are tuned are provided in Appendix 9.4.

Furthermore, I train a separate model for each of the six treatments using each machine learning method. These models are built on a subsample of the training data that includes only the control observations and individuals who received the respective treatment. The purpose of each model is to estimate the treatment effect for each specific treatment. As a result, I perform hyperparameter tuning individually for each of these six models within each method. Finally, the overall model assigns to each individual the treatment that has the highest estimated treatment effect according to these models.

5.2 Model Selection

A major challenge in tuning hyperparameters for models that predict treatment effects is the lack of a directly observable error term. Schuler et al. (2018) provide an overview of various metrics for assessing the performance of models that predict heterogeneous treatment effects. Based on their findings, they recommend using the metric $\widehat{\tau - \text{risk}}_R$ for model selection in individual treatment effect prediction. This metric originates from the R-Learner framework developed by Nie and Wager (2021), which employs the Robinson (1988) decomposition to reformulate the conditional average treatment effect in terms of the conditional mean outcome. For a more detailed explanation of the R-Learner method, please refer to Section 2.2.2. I also introduce the basic concept of the $\widehat{\tau - \text{risk}}_R$ metric in Appendix 9.3.

In their simulation study, Schuler et al. (2018) tested the $\widehat{\tau - \text{risk}}_R$ function and found that, when used for model selection, this particular loss function consistently chose models with a low mean squared error of the predicted treatment effect $E[(\hat{\tau}(X) - \tau(X))^2]$, outperforming other loss functions in terms of selecting accurate models.

The loss function used in the context of this thesis is defined as:

[7] K-fold cross-validation is a technique for evaluating predictive models. The dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time.

$$\widehat{\tau - \text{risk}}_R = \frac{1}{N_{te}} \sum_{i \in N_{te}}^{N_{te}} ((Y_i - \hat{m}(X_i)) - (T_i - \hat{p}(X_i))\hat{\tau}(X_i))^2, \tag{29}$$

where $\hat{m}(X_i)$ is an estimation of $E[Y_i|X_i]$ and $\hat{p}(X_i)$ is the estimation of the treatment propensity $E[T_i = 1|X_i]$.

To estimate $\hat{m}(X_i)$, I use a Lasso regression, and for propensity score $\hat{p}(x)$, I employ Logistic regression. Both models are cross-validated and trained on the training dataset, allowing them to predict $\hat{m}(X_i)$ and $\hat{p}(x)$ for the individuals in the test set. In addition, I apply the metric $\widehat{\tau - \text{risk}}_R$ as the model selection criterion for both the causal forest (CF) and causal neural networks (CNN) methods.

5.3 Comparison Strategy

For testing and comparing the performance of the two methods on an empirical dataset with unobserved true treatment effects, I use an approach based on the method outlined by Hitsch et al. (2024). The main idea of this approach is to use the observations in which coincidentally the predicted optimal treatment and the actual, randomly assigned, treatment are equal (matched). If the average outcomes of the matched observations are well above the mean treated outcome, this hints in the direction of models properly assigning optimal treatments.

First, the two models are trained on a training sample and predict the treatment effects for all respective treatments of the observations in the test sample. For each observation and prediction method in the test sample, the optimal treatment is assigned, according to the highest predicted treatment effect of the used treatments. For some observations, the assigned optimal treatment will be equal to the randomly assigned treatment. I refer to those observations as ***matched***.

The performance of the models can then be analyzed by examining the average outcome of the matched observations. For models that correctly assign the optimal treatment, the average outcome of matched observations will be higher than for models which do not. However, a high average outcome of matched observations does not necessarily coincide with the actual best treatment being assigned. It might be, e.g., the second-best (with a high treatment effect). Therefore, for most application examples using treatment assignments, correctly differentiating between effective and ineffective individual treatments is more important than predicting which one out of two very effective treatments is optimal.

Furthermore, if the mean outcome of matched observations is higher than the mean outcome of all treated observations, this suggests that assignment via the respective method is better than random assignment. If the mean outcome of matched observa-

tions is higher than the average outcome of individuals treated with a specific treatment, this suggests that assignment via the model is better than only assigning that specific treatment.

To ensure that the empirical results are not driven by randomness of the choice of the test and training samples, the described matched observation analysis (which I will refer to as ***Hitsch Matching***) are conducted via one-hundred times repeated three-fold cross-validation. Then, I calculate the average outcomes of matched treatments for each treatment and each method. In addition, I calculate the average outcomes of matched treatments overall (over all repetitions and folds).

Following the average points of the matched observations over all repetitions and folds, I will compare the distributions of the average outcomes of matched observations over the repetitions and the number/share of repetitions in which the average outcome of matched observations is higher than the average outcome of the best-performing treatment (Treatment 4), or higher than the average outcome over all treated individuals. As outlined in Section 4.2, the average points of treatments three and six are significantly below those of the other treatments. Therefore, I expected that the models will mostly predict much lower treatment effects for those treatments, such that they are almost never assigned as optimal. Consequently, there will be very few or no matched observations for treatments three and six. If a method is as good as random assignment for all treatments except three and six, and does not assign only treatments three and six as optimal, the average points of matched observations for the respective method will still be higher than the average points over all treatments.

To ensure that models do not perform well only because they mostly do not assign treatments three and six, I also compare the assignments only using treatments one, two, four, and five.

5.4 Motivation and Background to Winner's Curse

The *winner's curse* phenomenon in optimal treatment assignment, as explained by Andrews et al. (2024), underlines a bias where treatments with overestimated effects are more likely to be considered optimal. This phenomenon can lead to a systematic overestimation of the treatment effect for the selected optimal treatment. Consequently, the true optimal treatment may be overlooked because another treatment's effect was exaggerated. In this paper, I briefly introduce the winner's curse through a stylized example of average treatment effects. To solve the winner's curse issue, I propose shrinkage methods as a potential solution. Furthermore, in the next sections I present empirical findings using these shrinkage techniques.

5.4.1 A Stylized Example

Andrews et al. (2024) demonstrate the winner's curse using a simplified example of estimating average potential outcomes. Imagine there are n individuals randomly assigned to one of two groups: a treatment group where $T_i = 1$ and control group where $T_i = 0$, with $n/2$ individuals in each group. For each unit, we observe an outcome Y_i (such as the number of points scored). The average outcomes for the treatment and control groups, denoted as $Z_n^*(1)$ and $Z_n^*(0)$ respectively, are calculated as:

$$(Z_n^*(0), Z_n^*(1)) = \left(\frac{2}{n} \sum_{i=1}^{n} T_i Y_i, \frac{2}{n} \sum_{i=1}^{n} (1 - T_i) Y_i \right) \tag{30}$$

When participants are randomly sampled from the population (ensuring unconfoundedness), the estimators $Z_n^*(1)$ and $Z_n^*(0)$ become unbiased estimates of the population's average potential outcomes, represented as $(\mu^*(0), \mu^*(1)) = (E[Y_i^1], E[Y_i^0])$.

Given that the treatment is binary in this case, the set of possible policies is denoted by $\Theta = \{0, 1\}$. The optimal policy $\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} Z_n^*(\theta)$ is the one that yields the highest estimated average outcome.

Because there are only two possible policies in this example, determining the estimated optimal policy $\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} Z_n^*(\theta)$ becomes straightforward. Specifically, $\hat{\theta}$ becomes 1 if $Z_n^*(1) > Z_n^*(0)$, and equals 0 if $Z_n^*(1) < Z_n^*(0)$. Because the probability of a tie between $Z_n^*(1)$ and $Z_n^*(0)$ is zero, this decision rule is well-defined.

Assume that the estimated average outcomes for the control and treatment groups are jointly normally distributed with means $\mu(0), \mu(1)$, and variances $\Sigma(0), \Sigma(1)$. This can be expressed as:

$$\begin{pmatrix} Z(0) \\ Z(1) \end{pmatrix} \sim \mathbb{N} \left(\begin{pmatrix} \mu(0) \\ \mu(1) \end{pmatrix}, \begin{pmatrix} \Sigma(0) & 0 \\ 0 & \Sigma(1) \end{pmatrix} \right) \tag{31}$$

Given that the treatment is binary, the optimal policy $\hat{\theta}_n$ simplifies to choosing $\hat{\theta} = 1$ if $Z(1) > Z(0)$. Conditional on selecting this optimal policy ($\hat{\theta} = 1$) and observing a specific arbitrary value $Z(0) = z(0)$, the distribution of $Z(1)$ becomes a truncated normal distribution above $z(0)$. This truncation occurs because values of $Z(1)$ less than $z(0)$ would result in $\hat{\theta} \neq 1$, contradicting the initial condition $\hat{\theta} = 1$.

Because this truncation applies for all valid values of $z(0)$, it leads to a positive median bias in $Z(1)$ when conditioned on $\hat{\theta} = 1$. In other words, the median of $Z(1)$ exceeds its true mean $\theta(1)$ under this condition. Consequently, the probability that the estimated outcome for the chosen treatment exceeds its true average is greater than one-half:

$$P_\mu\{Z(\hat{\theta}) \geqslant \mu(\hat{\theta}) | \hat{\theta} = 1\} > \frac{1}{2}, \ \forall \mu \tag{32}$$

This inequality holds for all values of θ, indicating a systematic overestimation in the selected treatment's effect due to the winner's curse phenomenon.

Furthermore, this positive median bias applies symmetrically when the estimated optimal policy is $\hat{\theta} = 0$ and $Z(0) > Z(1)$. Consequently, the estimator $\hat{\theta}$ exhibits an unconditional positive median bias:

$$P_\mu\{Z(\hat{\theta}) \geqslant \mu(\hat{\theta})\} > \frac{1}{2}, \ \forall \mu \tag{33}$$

This implies that while $Z_n^*(\theta)$ is an unbiased estimator for $\mu^*(\theta)$ when policies are fixed, selecting policies based on these estimations leads to $Z_n^*(\hat{\theta}_n)$ systematically overestimating $\mu^*(\hat{\theta}_n)$. This stylized example can be extended to scenarios with more policies (e.g., additional treatments). Because every policy exhibits a conditional bias (as shown in Equation 32), the unconditional bias described in Equation 33 holds universally.

Moreover, while the previous example focused on average potential outcomes, the winner's curse phenomenon also applies to average treatment effects. Consider a scenario with two treatments (labeled 1 and 2) and a control group (labeled 0). The aim is to select the treatment that offers the highest estimated average treatment effect compared to the control. This selection is formalized as:

$$\hat{\theta}_n = \underset{\theta \in \{1,2\}}{\operatorname{argmax}} \left(Z_n^*(\theta) - Z_n^*(0) \right) \tag{34}$$

Conditional on choosing $\hat{\theta} = 1$ and the condition $Z(1) - Z(0) > Z(2) - Z(0)$, which simplifies to $Z(1) > Z(2)$, and given that $Z(2) = z(2)$, the distribution of $Z(1)$ becomes a normal distribution truncated below at $z(2)$. This truncation occurs because $Z(1)$ must exceed $Z(2)$ for treatment 1 to be selected. As a result, $Z(1)$ cannot take values less than $z(2)$ in this context. This leads to a persistent positive bias in the estimated effect of the selected treatment, demonstrating that the winner's curse still affects the estimation of average treatment effects.

When applying CATE as described in Section 2.1.2 to a specific individual i, we also encounter a truncation in the distribution of the estimated heterogeneous treatment effect when that treatment is chosen as optimal. This truncation arises because the treatment is selected based on higher estimated effects. The earlier stylized example illustrates that overestimation increases the likelihood of a treatment considered to be optimal. Whether the selection is based on the average outcome, average treatment effect, or CATE, the estimated effect of the optimal treatment tends to be overestimated rather than underestimated.

5.4.2 Shrinkage Estimators

As a potential way to address the winner's curse in optimal individual treatment assignment, I suggest applying shrinkage to the estimates, drawing them closer to a common mean. By doing so, predictions that were previously overestimated, and thus led to a particular treatment being chosen as optimal, are adjusted downward, because they are brought closer to the mean. This reduction in inflated predictions may result in assigning a different treatment that is genuinely more effective, thereby enhancing the estimator's performance in empirical analyses conducted earlier.

In the forthcoming discussion, I present two shrinkage estimators modified to suit the current problem of treatment effect prediction. The primary distinction between them lies in how they treat treatment estimates with high variance, specifically, whether such estimates should be shrunk more aggressively or less so.

James Stein Shrinker

The first shrinkage method I present is based on the James-Stein estimator introduced by Efron and Morris (1977). The Stein Paradox suggests that, even without considering covariates, using historical averages to estimate future averages, or the true underlying means, is not always the most effective approach. Specifically, their findings indicate that by adjusting these past averages toward a common overall mean, the estimators achieve a lower mean squared error and more accurately predict future averages in 16 of the 18 observations they studied.

The concept of this estimator can also be extended to heterogeneous treatment effects. Suppose we have estimates of the treatment effects $\hat{\tau}_i^k$ for each individual $i = 1, ..., n$ cross treatments $k = 1, ..., K$ (for example, obtained using the estimators mentioned earlier in Section 3.1.1). The shrinkage estimator for the heterogeneous treatment effect, denoted as $\hat{\phi}_{i,JS}^k$ is defined by:

$$\hat{\phi}_{i,JS}^k = \overline{\hat{\tau}^k} + c_{JS}^k(\hat{\tau}_i^k - \overline{\hat{\tau}^k}), \quad 1, ..., n \tag{35}$$

Here, $\overline{\hat{\tau}^k}$ represents the average estimated treatment effect for treatment k:

$$\overline{\hat{\tau}^k} = \frac{1}{n}\sum_{i}^{n} \hat{\tau}_i^k \tag{36}$$

The shrinkage factor c_{JS}^k for each treatment k is calculated as:

$$c_{JS}^k = 1 - \frac{(n-3)\sigma_{ATE}^2}{\sum_{i=1}^{n}(\hat{\tau}_i^k - \overline{\hat{\tau}^k})^2} \tag{37}$$

In this equation, σ_{ATE}^2 denotes the variance of the average treatment effect across

all six treatments. σ^2_{ATE} is determined by calculating the squared standard error of the coefficient β_1 from a simple OLS regression with a binary (dummy) variable. This regression utilizes the entire dataset and is specified as $y_i = \beta_0 + \beta_1 \tilde{T}_i + \varepsilon$, where $\tilde{T}_i$ is an indicator variable that equals 1 if the individual received any of the treatments being analyzed. The estimation is performed using the training set. Therefore, if the variance of the individual treatment effect estimates $\hat{\tau}^k_i$ is high relative to σ^2_{ATE}, the shrinkage factor c^k_{JS} approaches one, and the adjusted estimate $\hat{\phi}^k_{i,JS}$ moves closer to $\hat{\tau}^k_i$, indicating that the predictions of the model for the respective treatment are less shrunken. I refer to this method as *James-Stein Shrinker* (JS Shrinker).

Variance Shrinker

The second shrinkage method is inspired by the work of Chen and Zimmermann (2020). They apply a shrinkage estimator to correct for an upward bias in published stock returns reported in academic journals. This bias appears to result from journals favoring stock return predictors that generate large t-statistics, which in turn correspond to predictors with large sample mean returns. The shrinkage estimator, which I refer to as the *Variance Shrinker*, is defined by the following formula:

$$\hat{\phi}^k_{i,VS} = (1 - c^k_{VS})\hat{\tau}^k_i + c^k_{VS}\overline{\hat{\tau}^k} \tag{38}$$

where the shrinkage factor c^k_{VS} is given by:

$$c^k_{VS} = \frac{\sigma_k}{\sigma_{ATE} + \sigma_k} \tag{39}$$

In this equation, σ_{ATE} is the variance of the overall average treatment effect across all treatments, while σ_k is the variance of the average treatment effect specific to treatment k. The variance σ_k is measured by calculating the squared standard error of the coefficient β_1 from a simple OLS regression of the form $y_i = \beta_0 + \beta_1 \times T^k_i + \varepsilon$. In this regression, T^k_i is a binary indicator that equals 1 if individual i received treatment k, and 0 otherwise. The regression includes only observations from the control group and those who received treatment k, and it is performed using the training dataset.

As the variance σ_k increases while σ_{ATE} remains fixed, the shrinkage factor c^k_{VS} increases, approaching one. This means that the adjusted estimate $\hat{\phi}^k_{i,VS}$ is more heavily shrunk toward the mean $\overline{\hat{\tau}^k}$ of treatment k. In other words, treatments with higher variance in their average effects receive more shrinkage, which moves individual estimates closer to the treatment's overall mean prediction.

Shrinkage Variation Method

As previously described, both shrinkage methods adjust the predicted treatment effects by pulling them toward the mean of the predictions for each specific treatment, calculated as:

$$\overline{\hat{\tau}^k} = \frac{1}{n}\sum_{i}^{n} \hat{\tau}_i^k \tag{40}$$

However, if we suspect that simply receiving any one of the six (or four) treatments is the primary factor influencing outcomes, and that differences among these treatments are minimal, it may be more appropriate to shrink the estimates toward a common overall mean rather than treatment-specific means. To address this, I introduce modified versions of both shrinkage methods that adjust estimates toward the average treatment effect across all considered treatments.

The altered James-Stein Shrinker, which shrinks toward the overall mean, is defined as:

$$\hat{\phi}_{i,JS'}^k = \overline{\hat{\tau}} + c_{JS'}^k(\hat{\tau}_i^k - \overline{\hat{\tau}}), \quad 1, ..., n \tag{41}$$

where $\overline{\hat{\tau}}$ represents the average treatment effect calculated from the training set for all treatments:

$$\overline{\hat{\tau}} = \frac{1}{N(T_i = 1)}\sum_{i=1}^{n} T_i Y_i - \frac{1}{N(T_i = 0)}\sum_{i=1}^{n}(1 - T_i)Y_i \tag{42}$$

In this equation, T_i indicates whether individual i received any of the treatments under consideration. The shrinkage factor c^k for each treatment k is given by:

$$c_{JS'}^k = 1 - \frac{(n-3)\sigma_{ATE}^2}{\sum_{i=1}^{n}(\hat{\tau}_i^k - \overline{\hat{\tau}^k})^2} \tag{43}$$

Similarly, the modified Variance Shrinker is defined as:

$$\hat{\phi}_{i,JS'}^k = (1 - c_{VS'}^k)\overline{\hat{\tau}} + c_{VS'}^k \hat{\tau}_i^k, \tag{44}$$

with the shrinkage factor calculated by:

$$c_{VS'}^k = \frac{\sigma_{ATE}}{\sigma_{ATE} + \sigma_k} \tag{45}$$

Here, σ_{ATE} denotes the variance of the overall average treatment effect, and σ_k is the variance of the average treatment effect for treatment k. By shrinking toward the common overall mean, this modified method accounts for the possibility that the main

effect is due to being treated in general, rather than to differences between specific treatments.

Results

In this section, I report the results of the previously outlined empirical analysis. I describe the results when using all six treatments and subsequently the results of the subset of treatments, only using treatments one, two, four and five.

6.1 Results Without Shrinkage Estimators

Full Set of Treatments

Table 2 presents the average outcomes of matched observations for the full set of treatments, while Figure 8 illustrates the distribution of these averages over 100 cross-validation repetitions for the two machine learning methods. The causal forest (CF) model achieved an average of 1,981 points for matched observations, which is higher than the overall average across all treatments (1,898 points). This indicates that the CF model performs better in treatment assignment than does random allocation, making it the best-performing model in terms of average matched outcomes. In contrast, the causal neural network (CNN) model performed worse, with an average of 1,884 points for matched observations. This is below the overall average across all treatments (1,898 points), suggesting that the CNN model under performs random assignment.

	Treat 1	Treat 2	Treat 3	Treat 4	Treat 5	Treat 6	Overall
Causal Neural Network	1,931	1,932	1,766	1,967	1,912	1,863	1,884
	(18,162)	(23,384)	(24,762)	(11,879)	(1,746)	(5,661)	(85,594)
Causal Forest	2,141	1,899	1,927	1,785	2,189	2,124	1,981
	(7,731)	(6,062)	(26)	(37,786)	(30,229)	(2,308)	(84,142)
Average	1,926	1,930	1,764	1,970	1,931	1,871	1,898
	(87,900)	(86,500)	(87,500)	(84,800)	(87,400)	(84,500)	(518,600)

Table 2: *Mean Outcome of Matched Observations: Full Treatment Set and No Shrinkage Applied*: This table shows the results of the three-fold cross-validation of the Hitsch Matching repeated 100 times. All six treatments were considered. For each of the two machine learning methods, the table predicts the average outcome of matched observations over all folds and repetitions and in brackets the number of observations that were matched in total. This is shown for each treatment and all treatments. The last row depicts the average points for the participants in the respective treatments and the overall average.

Notably, the CF model average matched outcome (1,981 points) is also higher than the average outcome of treatment 4 (1,970 points). However, for both models, the average outcome of matched observations specifically within treatment 4 is lower than the average outcome of that treatment. For the CF model, the average matched outcome within treatment 4 is 1,785 points, which is significantly below 1,970 points.

Furthermore, the CF model assigns treatment 4 most frequently, which is expected given that treatment 4 has the highest overall average outcome. This frequent assignment could be due to the model systematically overestimating the treatment effects for treatment 4, or because treatment 4 serves as a "baseline" treatment within the models. If the predicted treatment effects for other treatments are low, perhaps because the models do not detect significant potential in them, the prediction for treatment 4 remains relatively high, leading to its frequent assignment.

An exceptional result is observed for treatment 5 in the CF model, where the matched average outcome is 2,189 points compared to the treatment's average of 1,931 points. This suggests that the model effectively assigns participants who are more competitively oriented to the "Real-Time Feedback" treatment.

When measuring the number of repetitions where the average matched outcome exceeded the average outcome of treatment 4, the CF model significantly outperforms the CNN model. The CF model achieved this in 79 of 100 repetitions, whereas the CNN model did so in only 1 of 100 repetitions. Additionally, the CNN model frequently had overall average matched outcomes worse than the mean outcome of treated observations, occurring in 61 of 100 repetitions.

Focusing on the CF model alone, the results indicate that assigning treatments using this model is more effective than assigning everyone to the best-performing treatment. According to the metrics introduced, the CNN neural network-based model performs much worse than the tree-based CF model.

Subset of Treatments

Previously, a challenge in the comparison method arose because treatments 3 and 6 had outcomes significantly lower than the other treatments. This discrepancy could cause models to perform well merely by avoiding assigning treatments 3 and 6 as optimal, therefore skewing the comparison results. To address this issue, I conduct an additional analysis using only a subset of treatments, specifically treatments 1, 2, 4, and 5.

The results for this subset are presented in Table 3, with the distributions across repetitions illustrated in Figure 9. The causal forest (CF) model continues to outperform the causal neural network (CNN) model. The CF model achieved an average outcome of matched observations of 1,983 points (up from 1,981 previously), with 99 of 100 repetitions exceeding the mean outcome (now 1,939 when only the subset of four treatments are considered). Moreover, the number of repetitions where the average outcome surpassed that of treatment 4 increased significantly, to 88 of 100.

In contrast, the metrics for the CNN model remain similar to those when all six treatments are used, and are considerably lower than those of the CF model. The CNN

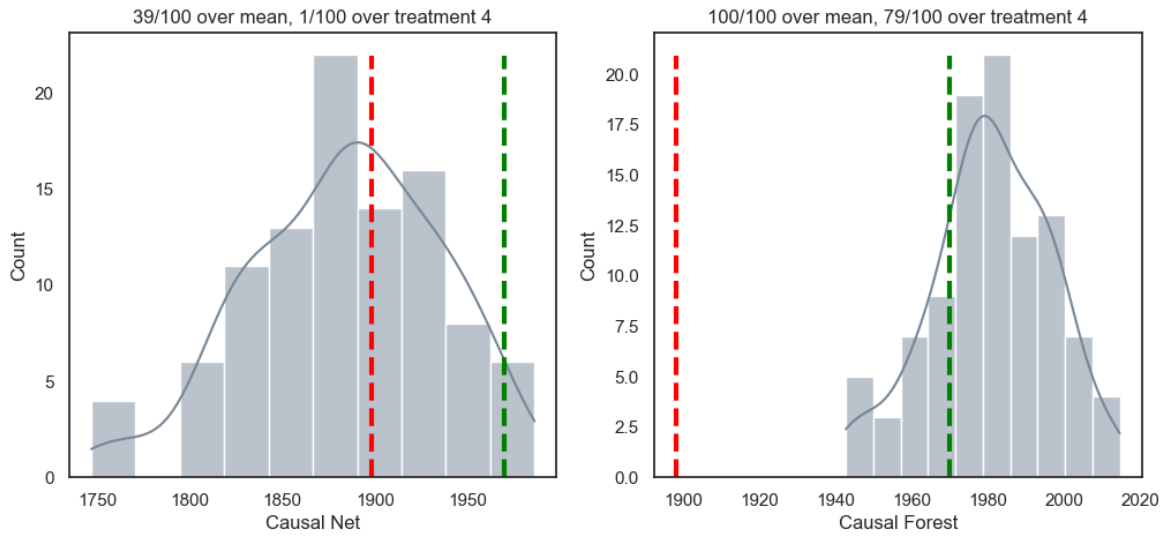

Figure 8: *Hitsch Matching - Full Treatment Set*: This figure depicts the distribution of the average outcome of matched observations of the individual 100 repetitions of the three-fold cross-validation. All six treatments were considered. The green line depicts the average outcome of participants treated with treatment four (loss treatment), 1,970 points, and the red line depicts the average outcome of participants treated with any of the treatments, 1,898 points.

model's average outcome of matched observations over the hundred repetitions is 1,939 points, higher than in the full sample (1,884 points) but still below the average outcome of treatment 4 (1,970 points) and equal to the mean outcome of the four treatments considered. As with the previous results, the CNN model's average outcome exceeds that of treatment 4 in only 4 of 100 cross-validation repetitions, and in 52 of 100 repetitions, it falls below the mean outcome of the considered subset of four treatments.

Overall, because the results using the subset of treatments are quite similar to those obtained with the full treatment set, there is no substantial concern about the validity of the findings of this paper. This suggests that using treatments with relatively uniform average outcome levels does not significantly affect the conclusions compared to using treatments with varying average outcomes.

6.2 Results With Shrinkage Estimators

As in Section 6.1, I start by presenting the results that include all six treatments, and then proceed to discuss the results for the subset of treatments.

Full Set of Treatments

Using all six treatments, Table 4 presents the average outcomes of matched observations for two estimation methods combined with the applied shrinkage techniques. Figures

	Treat 1	Treat 2	Treat 4	Treat 5	Overall
Causal Neural Network	1,933	1,928	1,972	1,915	1,939
	(30,808)	(32,321)	(19,045)	(2,651)	(84,825)
Causal Forest	2,161	1,918	1,789	2,185	1,983
	(8,797)	(6,583)	(38,092)	(30,762)	(84,234)
Average	1,926	1,930	1,970	1,931	1,939
	(87,900)	(86,500)	(84,800)	(87,400)	(346,600)

Table 3: *Mean Outcome of Matched Observations: Subset of Treatments and No Shrinkage Applied*: This table shows the results of the three-fold cross-validation of the Hitsch Matching repeated 100 times. Only treatments 1, 2, 4, and 5 were considered. For each of the two machine learning methods, the table depicts the average outcome of matched observations over all folds and repetitions, and, in brackets the number of observations that were matched in total. This is shown for each treatment and all treatments. The last row depicts the average points for the participants in the respective treatments and the overall average.

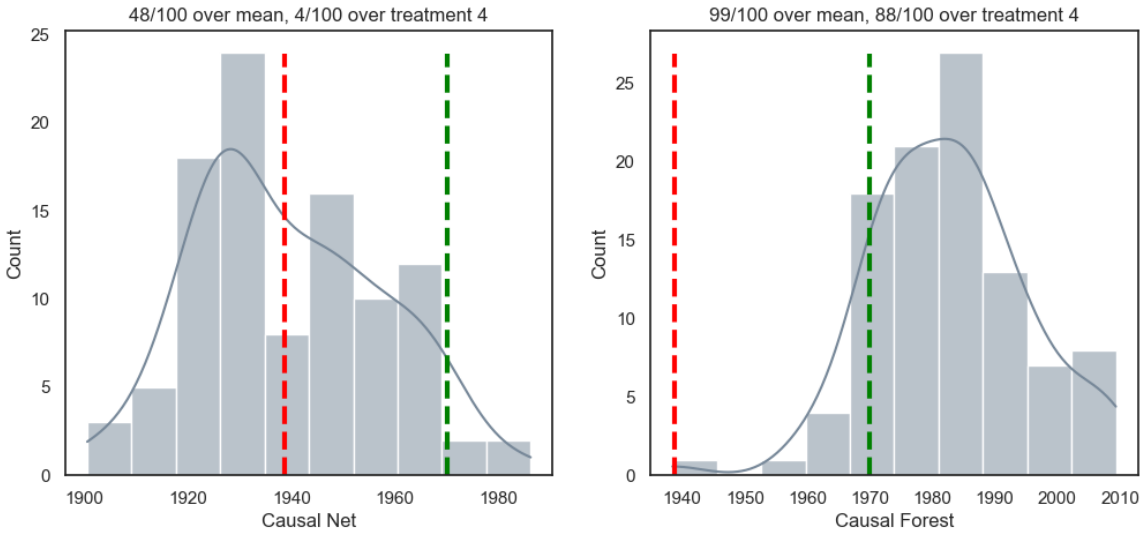

Figure 9: *Hitsch Matching - Subset of Treatments*: This figure depicts the distribution of the average outcome of matched observations of the 100 individual repetitions of the three-fold cross-validation. Only treatments 1, 2, 4, and 5 were considered. The green line depicts the average outcome of participants treated with treatment four (loss treatment), 1,970 points, and the red line depicts the average outcome of participants treated with any of the four considered treatments, 1,939 points.

10 and 11 illustrate the distributions of these methods when used alongside the four different shrinkage methods. Specifically, when applying the James-Stein Shrinker that adjusts estimates toward the average treatment effect prediction of each respective treatment, the average outcome for the causal forest (CF) method increases slightly

from CF $= 1,981$ (the baseline without shrinkage) to CF $\overline{\hat{\tau}_k} = 1,988$. Additionally, the number of repetitions where the causal forest's matched observations exceed the average outcome of treatment 4 rises from 79 to 88 of 100. This marks the best overall performance among all specifications and methods in this empirical analysis. Although the performance metrics for the causal neural network (CNN) also improve slightly with this shrinkage method, its overall performance remains low.

When using the James-Stein Shrinker, which adjusts estimates toward the overall average treatment effect across all six treatments, the performance of the CF remains largely unaffected, but there is an improvement for the CNN. However, for the CF, there are no repetitions in which the average points of matched observations surpass the average outcome of treatment 4. The CF's matched observations exceed the overall average outcome of all six treatments in only 54 of 100 repetitions and surpass the average outcome of treatment 4 in just 2 of 100 repetitions. These results indicate that, among all combinations of methods and shrinkage techniques, this particular shrinker applied to the CNN predictions performs the worst. Furthermore, the average points of matched observations for CNN drop to CNN JS $\overline{\hat{\tau}_k} = 1,892$, which is lower than the average outcomes of all treatments except for treatments 3 and 6.

Applying the Variance Shrinker that adjusts estimates toward the average predicted treatment effect of each respective treatment decreases the CF's average points of matched observations from CF $= 1,981$ to CF Var $\overline{\hat{\tau}_k} = 1,975$. The number of repetitions where matched observations exceed the average outcome of treatment 4 also decreases from 79 to 70 of 100. For the CNN, both performance metrics remain almost unchanged, continuing to exhibit very low performance.

Using the Variance Shrinker that adjusts estimates toward the overall average treatment effect results in a decrease in both the average points of matched observations and the number of repetitions in which these points exceed the overall average outcome and the average outcome of treatment 4. This decline is observed across all estimators, with performance falling well below the initial values of the baseline methods.

Overall, the shrinkage methods have varying effects on the two machine learning approaches. CNN's already poor performance could not be meaningfully improved and even deteriorated further with the application of shrinkage. In contrast, the tree-based method of CF experienced performance improvements when shrinkage techniques were applied. This improvement may be because, for tree-based methods, the winner's curse was, or still is, a problem hindering optimal treatment assignment, and shrinkage helps mitigate this issue. For the CNN, however, due to its overall poor performance, the winner's curse is likely not the primary factor limiting its effectiveness. Shrinkage methods that adjust estimates toward the overall mean generally decreased performance metrics for both methods and both shrinkage techniques, with few instances where the

shrinkage was beneficial. This may be because the differences between treatments were so significant that adjusting estimates toward a common overall mean resulted in excessive shrinkage.

Subset of Treatments

The results using the subset of treatments, specifically treatments 1, 2, 4, and 5, are presented in Table 5. Figures 12 and 13 display the distributions of the average outcomes across repetitions.

As discussed in Section 5.4.2, if the true underlying treatment effects are similar across treatments, shrinkage methods that adjust estimates toward the overall average treatment effect (ATE) are expected to perform better than those shrinking toward treatment-specific average predicted effects. In the case of using all six treatments, as outlined in Section 4.2, this assumption is unlikely. This likely explains why shrinkers targeting the overall average performed worse than both the baseline methods, and why the predictions shrank toward individual treatment averages. Consistent with this reasoning, when analyzing only treatments 1, 2, 4, and 5, which presumably have more similar treatment effects, the shrinkers pulling toward the overall ATE performed much better.

Focusing on the causal forest (CF) method, the James-Stein Shrinker (JS-Shrinker) provided significant performance improvements, similar to the results with the full treatment set. While the JS-Shrinker that shrinks toward the overall ATE performed poorly with all six treatments, it improved the average outcome of matched observations from CF $= 1,983$ to CF JS $\overline{\hat{\tau}} = 1,987$ in the subset case. Additionally, the number of repetitions in which matched observations exceeded the average points of treatment 4 increased from 88 to 92 of 100. When shrinking toward the average predicted treatment effect of each treatment, the improvements were nearly identical, with CF JS $\overline{\hat{\tau}_k} = 1,986$ and 88 of 100 repetitions, respectively.

In line with the findings using the full treatment set, the Variance Shrinker did not substantially improve the CF's predictions. The performance metrics for the shrinker targeting the overall mean remained almost unchanged. However, when shrinking toward the treatment-specific mean, the average points of matched observations decreased to 1,974, and the number of repetitions where the average points exceeded those of treatment 4 dropped to 63 of 100.

For the causal neural network (CNN) method, the shrinkers provided only minimal performance improvements. The shrinkers that slightly increased performance from the baseline CNN $= 1939$ were CNN JS $\overline{\hat{\tau}_k} = 1,940$, CNN Var $\overline{\hat{\tau}_k} = 1,940$

and CNN Var $\overline{\hat{\tau}} = 1947$. However, for most shrinker variations, the average points of matched observations decreased and remained low overall.

Although the JS-Shrinker shrinking toward the overall mean of the four treatments increased the number of repetitions in which matched observations exceeded the average outcome of treatment 4 from 4 to 6 of 100 times, it also increased the number of repetitions in which the average points fell below the average of all four treatments from 52 to 58 of 100 times. This may happen because using this shrinker causes greater variation in the average points of matched observations between repetitions compared to applying other shrinkers to the CNN predictions (see Figure 13).

In summary, shrinkers that adjust estimates toward the overall ATE performed significantly better when applied to treatments with similar outcome levels. In the subset of treatments, the James-Stein Shrinker outperformed the Variance Shrinker for the tree-based CF method. Conversely, when the full treatment set is analyzed, the relative performance between the shrinkers was more vague.

Discussion

In this section I summarize the results and the few limitations of this paper. **First**, when comparing the causal neural network (CNN) and the causal forest (CF) with the Hitsch Matching method on an empirical dataset, I find that the CF performs best in assigning the optimal treatment. The findings indicate that selecting the optimal treatment based on the highest predicted treatment effect from the CF yields a higher outcome level than simply assigning the best overall performing treatment. In contrast, the CNN only performs marginally better than random treatment assignment.

The CNN performs significantly worse than the CF, and several factors may explain this disparity in the performance. First, a key factor for the weak performance of CNN, also a key limitation of this study, is the sample size. Approximately 1,160 observations are available for training each neural network, which may be insufficient, especially since neural networks necessitate splitting data into validation samples. Specifically, there are about 864 observations per treatment and 879 for the control group, totalling around 1,740 observations per network. With 2/3 allocated to the training set ($1,740 \times \frac{2}{3} = 1,160$), this amount of data may be inadequate for the neural network to learn complex patterns effectively. To overcome this limitation, follow-up research could focus on selecting a use case with a larger dataset that will be more conducive to neural network training. This would enable us to perform a more rigorous benchmarking of the CF and CNN methods, and allow for more accurate conclusions about their relative performance.

Second, tuning neural networks is challenging due to the high interdependence of hyperparameters compared to the tree-based CF. The hyperparameter optimization and the error proxy ($\tau - \text{risk}_R$) used in this study may not have been suitable for the CNN. To address this issue, a potential solution would be to test the models using a different meta-learner instead of the *R-learner* ($\tau - \text{risk}_R$). Employing an alternative meta-learner may provide a better fit for the neural network architecture and improve the hyperparameter tuning process.

Third, the selected hyperparameter grids might have been less than optimal for this dataset and could have been expanded, but doing so was computationally infeasible for this study. However, a potential solution could be to use another hyperparameter optimization framework, such as Optuna. Employing Optuna can speed up the hyperparameter tunning process because it efficiently searches large spaces and prunes unpromising trials for faster results.

Second, I address the issue of the winner's curse, in which optimal treatments are systematically overestimated, which is a significant challenge in optimal treatment assignments using predicted HTEs. To mitigate this issue, I introduce the James-Stein

and Variance shrinkers as a solution. In addition, I introduce a modified version of these techniques that adjusts estimates toward the average treatment effect across all considered treatments instead. Applying these shrinkage methods improved prediction performance in some cases. The shrinkers' effectiveness varied across models and treatment subsets: shrinkers that adjusted towards the overall average outcome performed poorly when using all six treatments, but much better with the subset of four similar treatments. Overall, the James-Stein Shrinker led to greater performance improvements than the Variance Shrinker. The tree-based CF method benefited the most from applying shrinkers, whereas CNN's performance did not show notable improvements.

The differences between the machine learning (ML) methods and shrinkage techniques are small, and the effects depend heavily on the specific split used in cross-validation. Several factors could explain these small differences. For instance, although Amazon Mechanical Turk (MTurk) has advantages over traditional experimental setups, as discussed in Section 2, the dataset is likely noisier, and participants may have been less sincere when completing the surveys compared to a subset of individuals who would typically be better compensated and may have more interest in the research. For example, studies conducted in dedicated research facilities with professional participants or specialized panels, where participants receive higher compensation and are vested in the research, can result in more reliable data.

Moreover, as previously discussed, the dataset may be small to adequately capture the effects or to allow the ML models to be appropriately trained and compared. In addition, another possibility is that the selected treatments and experimental setup do not allow for significant variation in heterogeneous treatment effects; however, findings from Opitz et al. (2024) suggest otherwise.

Lastly, I recommend exploring and understanding the mechanisms behind the improvements in predictions when using shrinkage techniques as follow-up research. Although shrinkage estimators improve the optimal treatment assignment of the models, the underlying mechanisms remain unclear and have not been extensively explored in the literature. This study's shrinkage methods are adapted from classical shrinkage theories and concepts. They are not based on established asymptotic properties in the context of estimating treatment effects for optimal treatment assignment. Understanding these mechanisms is crucial, given that shrinkers perform differently across methods and treatment sets. Additionally, the shrinkers used in this paper are adapted from existing shrinkage concepts and are not based on proven asymptotic properties in estimating treatment effects for optimal treatment assignment. By clarifying these mechanisms, it would be possible to develop optimal shrinkers specifically designed to address the winner's curse in optimal treatment assignment using ML methods.

Conclusion

This paper compares two machine learning (ML) techniques, the *direct* estimation method of Causal Forests (CF) and the *indirect* method of Causal Neural Networks (CNN), for predicting heterogeneous treatment effects (HTEs) for optimal treatment assignment in a setting where multiple treatments are available. To empirically compare the two methods, I use a dataset from an online experiment on incentivizing manual labour Opitz et al. (2024), using a comparison strategy based on Hitsch et al. (2024) and Nie and Wager (2021). First, I conducted an extensive literature review that addresses critical challenges in estimating conditional average treatment effects (CATE). Then, I introduced the problem of optimal treatment assignment and discussed meta-learner techniques for estimating CATE. Additionally, I explored popular machine learning methods commonly employed by economists in causal inference, focusing on optimal treatment assignment.

By comparing the CNN and CF methods using the *Hitsch Matching* approach on the empirical dataset, I find that the CF method achieves the best performance in assigning individuals the optimal treatment from a given set of treatments. The results indicate that selecting treatments based on the highest predicted treatment effect from the CF leads to higher outcome levels than simply assigning the overall best-performing treatment to all individuals. In contrast, the CNN method performs only marginally better than random treatment assignment and significantly worse than the CF method.

Moreover, I address the issue of the *winner's curse*, where optimal treatments are systematically overestimated, which is a significant challenge in optimal treatment assignment using predicted HTEs. To overcome this issue, I introduce two families of shrinkage estimators: the James-Stein shrinker and the Variance shrinker. This paper contributes to the ML literature for optimal targeting by applying these techniques in a novel context: estimating treatment effects for optimal treatment assignment.

I find that employing shrinkage methods can enhance the performance of predictions in most cases. The shrinkage estimators performed differently across the two models and the subsets of treatments. Notably, shrinkers that adjust predictions toward the overall average outcome performed less accurately when analyzing all six treatments but showed improved performance when focusing on a subset of four similar treatments. Overall, the James-Stein Shrinker resulted in greater performance improvements compared to the Variance Shrinker. In addition, the CF method benefited significantly from applying shrinkage estimators, whereas the CNN method did not show notable improvements. However, the differences between the performance of the shrinkage estimators for the two ML methods are relatively small and are heavily dependent on the cross-validation split. Therefore, we should be cautious about drawing definitive

conclusions from the empirical results.

Finally, I intend this paper to provide a guideline for systematically selecting and comparing different estimation methods to predict optimal targeting policies. Studying methods for optimal treatment assignment is crucial because it enables targeted policies that maximize the effectiveness of treatments while minimizing potential harm. By leveraging ML for causal inference, researchers can predict individual responses to treatments more accurately, and ensure that resources are allocated to those who will benefit most. Optimal treatment assignment can thus enhance processes in fields like healthcare, education, and other policymaking by offering personalized treatments and improving overall outcomes.

References

Monica Andini, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio, and Viola Salvestrini. Targeting with machine learning: An application to a tax rebate program in italy. *Journal of Economic Behavior & Organization*, 156:86–102, 2018.

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on winners. *The Quarterly Journal of Economics*, 139(1):305–358, 2024.

Joshua D Angrist and Alan B Krueger. Empirical strategies in labor economics. In *Handbook of labor economics*, volume 3, pages 1277–1366. Elsevier, 1999.

Timothy Armstrong and Shu Shen. Inference on optimal treatment assignments. 2015.

Eva Ascarza. Retention futility: Targeting high-risk customers might be ineffective. *Journal of marketing Research*, 55(1):80–98, 2018.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic perspectives*, 31(2):3–32, 2017.

Susan Athey and Stefan Wager. Policy learning with observational data, 2020. URL `https://arxiv.org/abs/1702.02896`.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. 2019.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Andrew Y Chen and Tom Zimmermann. Publication bias and the cross-section of stock returns. *The Review of Asset Pricing Studies*, 10(2):249–289, 2020.

Shuai Chen, Lu Tian, Tianxi Cai, and Menggang Yu. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 73(4): 1199–1209, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Stefano DellaVigna and Devin Pope. What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069, 2018.

Bradley Efron and Carl Morris. Stein's paradox in statistics. *Scientific American*, 236 (5):119–127, 1977.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Carlos Fernández-Loría, Foster Provost, Jesse Anderton, Benjamin Carterette, and Praveen Chandar. A comparison of methods for treatment assignment with an application to playlist generation. *Information Systems Research*, 34(2):786–803, 2023.

D Jake Follmer, Rayne A Sperling, and Hoi K Suen. The role of mturk in education research: Advantages, issues, and future directions. *Educational Researcher*, 46(6): 329–334, 2017.

Xiajing Gong, Meng Hu, Mahashweta Basu, and Liang Zhao. Heterogeneous treatment effect analysis based on machine-learning methodology. *CPT: Pharmacometrics & Systems Pharmacology*, 10(11):1433–1443, 2021.

Keisuke Hirano and Jack R Porter. Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701, 2009.

Günter J Hitsch, Sanjog Misra, and Walter W Zhang. Heterogeneous treatment effects and optimal targeting policy evaluation. *Quantitative Marketing and Economics*, 22 (2):115–168, 2024.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

Torben SD Johansen. Optimal treatment allocation under constraints. *arXiv preprint arXiv:2404.18268*, 2024.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

Oliver P John. The big five inventory—versions 4a and 54. *(No Title)*, 1991.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–495, 2015.

Noémi Kreif, Richard Grieve, Rosalba Radice, and Jasjeet S Sekhon. Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. *Health Services and Outcomes Research Methodology*, 13:174–202, 2013.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Alexander R Luedtke and Mark J Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016.

Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.

Charles F Manski and Aleksey Tetenov. Trial size for near-optimal choice between surveillance and aggressive treatment: Reconsidering mslt-ii. *The American Statistician*, 73(sup1):305–311, 2019.

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51, 1923.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Saskia Opitz, Dirk Sliwka, Timo Vogelsang, and Tom Zimmermann. The algorithmic assignment of incentive schemes. *Management Science*, 2024.

Judea Pearl. Causal inference in statistics: An overview. 2009.

James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: analysis of correlated data*, pages 189–326. Springer, 2004.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

Duncan Simester, Artem Timoshenko, and Spyros I Zoumpoulis. Efficiently evaluating targeting policies: Improving on champion vs. challenger experiments. *Management Science*, 66(8):3412–3424, 2020.

Anthony Strittmatter. What is the value added by using causal machine learning methods in a welfare experiment evaluation? *Labour Economics*, 84:102412, 2023.

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672, 2016.

Aleksey Tetenov. Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, 166(1):157–165, 2012.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113 (523):1228–1242, 2018.

Richard J Willke, Zhiyuan Zheng, Prasun Subedi, Rikard Althin, and C Daniel Mullins. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC medical research methodology*, 12:1–12, 2012.

Jeffrey M Wooldridge. Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420–445, 2015.

Qi Xu, Haoda Fu, and Annie Qu. Optimal individualized treatment rule for combination treatments under budget constraints. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad141, 2024.

Appendix

9.1 Figures and Tables

	Treat 1	Treat 2	Treat 3	Treat 4	Treat 5	Treat 6	Overall
CNN	1,931	1,932	1,766	1,967	1,912	1,863	1,884
	(18,162)	(23,384)	(24,762)	(11,879)	(1,746)	(5,661)	(85,594)
CNN JS $\overline{\hat{\tau}_k}$	1,925	1,935	1,761	1,970	1,912	1,875	1,889
	(19,699)	(21,875)	(21,155)	(12,685)	(1,746)	(5,847)	(83,007)
CNN JS $\overline{\hat{\tau}}$	1,933	1,929	1,764	1,973	1,932	1,865	1,892
	(23,052)	(24,090)	(19,570)	(9,016)	(3,159)	(6,475)	(85,362)
CNN Var $\overline{\hat{\tau}_k}$	1,926	1,942	1,769	1,968	1,951	1,845	1,994
	(10,588)	(6,352)	(4,051)	(23,676)	(25,915)	(15,443)	(86,015)
CNN Var $\overline{\hat{\tau}}$	1,936	1,953	1,731	1,969	2,004	1,799	1,891
	(9,259)	(4,489)	(12,599)	(12,103)	(22,724)	(23,964)	(85,126)
CF	2,141	1,899	1,927	1,785	2,189	2,124	1,981
	(7,731)	(6,062)	(26)	(37,786)	(30,229)	(2,308)	(84,142)
CF JS $\overline{\hat{\tau}_k}$	2,163	1,967	2,768	1,794	2,194	2,144	1,988
	(6,808)	(5,174)	(1)	(39,847)	(31,654)	(684)	(84,168)
CF JS $\overline{\hat{\tau}}$	2,141	1,919	1,964	1,775	2,180	2,151	1,988
	(5,542)	(4,989)	(30)	(35,939)	(35,486)	(2,116)	(84,102)
CF Var $\overline{\hat{\tau}_k}$	2,145	1,997	-	1,842	2,221	2,122	1,975
	(7,764)	(4,549)	(0)	(50,206)	(21,264)	(154)	(83,937)
CF Var $\overline{\hat{\tau}}$	2,154	1,908	1,922	1,789	2,194	2,137	1,982
	(7,585)	(6,018)	(24)	(38,632)	(29,644)	(2,263)	(84,166)
Average	1,926	1,930	1,764	1,970	1,931	1,871	1,898
	(87,900)	(86,500)	(87,500)	(84,800)	(87,400)	(84,500)	(518,600)

Table 4: *Mean Outcome of Matched Observations: Full Treatment Set and Shrinkage Applied*
The table presents the outcomes from a three-fold cross-validation of the Hitsch Matching method, repeated 100 times. Here, I analyze all six treatments. The table shows the average results of the matched observations across all folds and repetitions for each machine learning method combined with the four shrinkage techniques introduced earlier. I provide the total number of matched observations in parentheses. These findings are displayed for each treatment and collectively for all treatments. The final row lists the average scores for participants within each treatment group and the overall average across all groups. In the table, "CNN" stands for Causal Neural Network, "CF" for Causal Forest, "JS" for James-Stein Shrinker and "Var" for Variance Shrinker. The notation $\overline{\hat{\tau}_k}$ indicates that the shrinkage methods adjust towards the average treatment prediction specific to each treatment, while $\overline{\hat{\tau}}$ denotes the shrinkage towards the overall average treatment effect across all treatments included in the analysis.

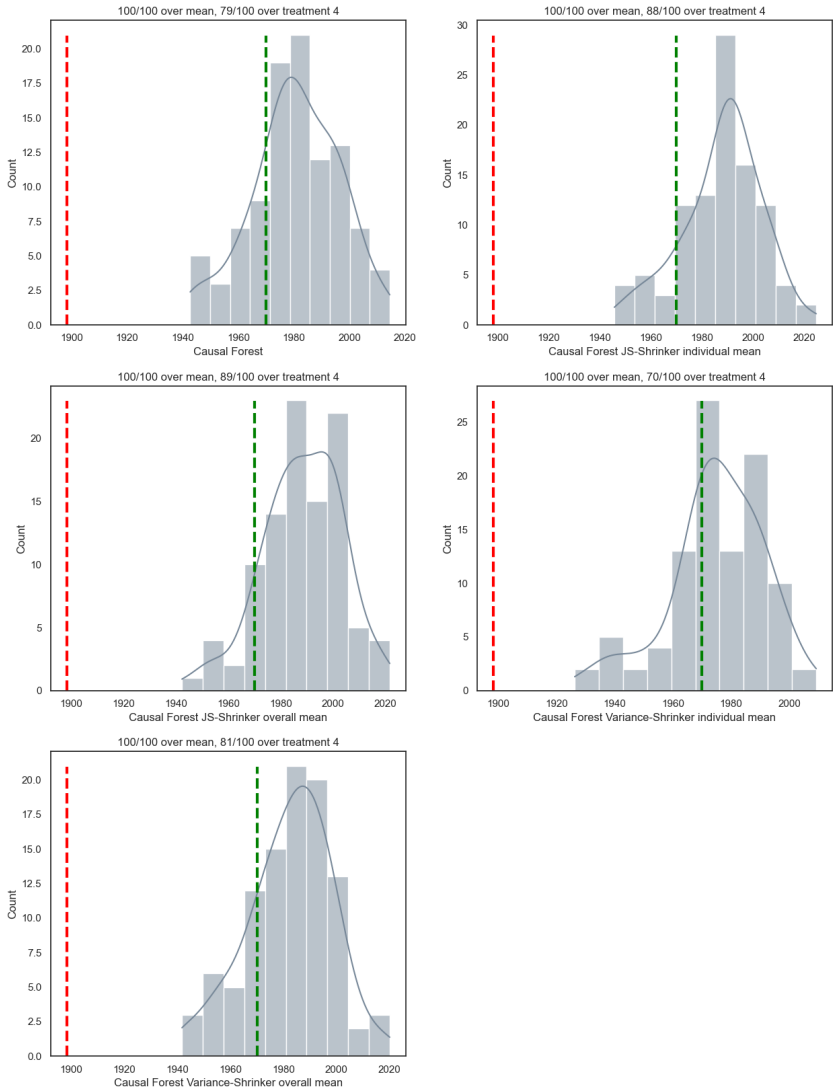

Figure 10: *Hitsch Matching: Using CF, Shrinkage Estimators, and the Full Set of Treatments*

The figure shows the distribution of average outcomes from matched observations across 100 individual repetitions of 3-fold cross-validation, using Causal Forests and all four shrinkage methods. All six treatments are included in the analysis. The green line represents the average outcome of participants who received treatment four (the loss treatment), equal to 1,970 points. The red line indicates the average outcome of participants who received any of the treatments, equal to 1,898 points. The label "individual mean" refers to shrinkage methods that adjust predictions toward the mean predicted treatment effect of each specific treatment. Conversely, "overall mean" refers to shrinkage methods that adjust predictions toward the average treatment effect of receiving any of the treatments used (as calculated in the training set).

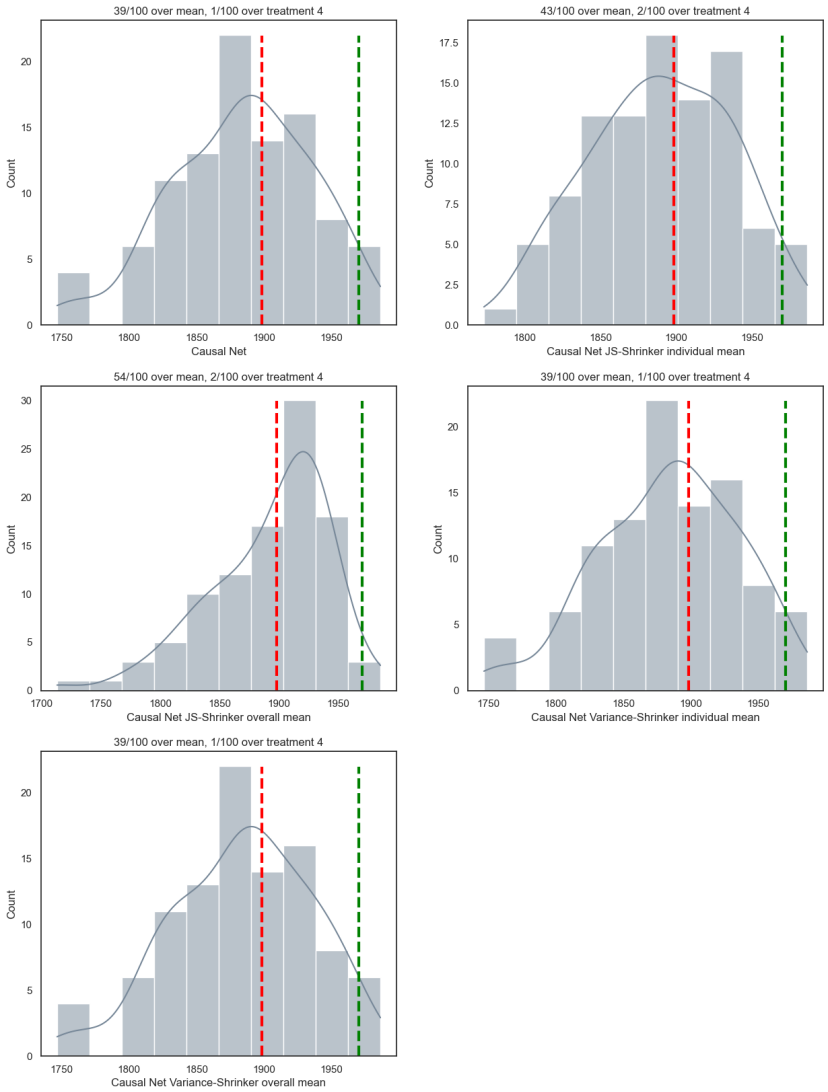

Figure 11: *Hitsch Matching: Using CNN, Shrinkage Estimators, and the Full Set of Treatments*

The figure shows the distribution of average outcomes from matched observations across 100 individual repetitions of 3-fold cross-validation, using Causal Neural Network and all four shrinkage methods. All six treatments are included in the analysis. The green line represents the average outcome of participants who received treatment four (the loss treatment), equal to 1,970 points. The red line indicates the average outcome of participants who received any of the treatments, equal to 1,898 points. The label "individual mean" refers to shrinkage methods that adjust predictions toward the mean predicted treatment effect of each specific treatment. Conversely, "overall mean" refers to shrinkage methods that adjust predictions toward the average treatment effect of receiving any of the treatments used (as calculated in the training set).

	Treat 1	Treat 2	Treat 4	Treat 5	Overall
CNN	1,933	1,928	1,972	1,915	1,939
	(30,808)	(32,321)	(19,045)	(2,651)	(84,825)
CNN JS $\overline{\hat{\tau}_k}$	1,929	1,933	1,970	1,934	1,940
	(28,507)	(28,160)	(17,719)	(3,218)	(77,604)
CNN JS $\overline{\hat{\tau}}$	1,929	1,928	1,975	1,930	1,937
	(39,136)	(27,512)	(15,351)	(2,679)	(84,678)
CNN Var $\overline{\hat{\tau}_k}$	1,922	1,930	1,967	1,943	1,946
	(13,844)	(9,688)	(29,075)	(33,628)	(86,223)
CNN Var $\overline{\hat{\tau}}$	1,908	1,934	1,955	1,961	1,947
	(14,455)	(9,414)	(29,883)	(32,645)	(86,385)
CF	2,161	1,918	1,789	2,185	1,983
	(8,797)	(6,583)	(38,092)	(30,762)	(84,234)
CF JS $\overline{\hat{\tau}_k}$	2,168	1,975	1,794	2,194	1,987
	(7,282)	(5,447)	(40,006)	(31,388)	(84,123)
CF JS $\overline{\hat{\tau}}$	2,158	1,949	1,780	2,185	1,987
	(7,899)	(8,545)	(35,722)	(32,121)	(84,287)
CF Var $\overline{\hat{\tau}_k}$	2,137	1,998	1,844	2,221	1,974
	(8,175)	(4,567)	(50,457)	(20,695)	(83,894)
CF Var $\overline{\hat{\tau}}$	2,162	1,921	1,791	2,187	1,983
	(8,736)	(6,524)	(38,509)	(30,454)	(84,223)
Average	1,926	1,930	1,970	1,931	1,939
	(87,900)	(86,500)	(84,800)	(87,400)	(346,600)

Table 5: *Mean Outcome of Matched Observations: Subset of Treatment Set and Shrinkage Applied*
The table summarizes the outcomes from 100 repetitions of three-fold cross-validation using the Hitsch Matching method, considering only treatments 1, 2, 4, and 5. The table displays the average outcome of the matched observations across all folds and repetitions for each machine learning algorithm, combined with the four shrinkage methods introduced earlier. The total number of matched observations is indicated in parentheses next to each result. These outcomes are presented for each treatment and collectively for all treatments. The final row shows the average scores achieved by participants in each specific treatment and the overall average across all treatments. In the table, "CNN" stands for Causal Neural Network, "CF" for Causal Forest, "JS" for James-Stein Shrinker and "Var" for Variance Shrinker. The notation $\overline{\hat{\tau}_k}$ indicates that the shrinkage methods adjust towards the average treatment prediction specific to each treatment, while $\overline{\hat{\tau}}$ denotes shrinkage towards the overall average treatment effect across all treatments included in the analysis.

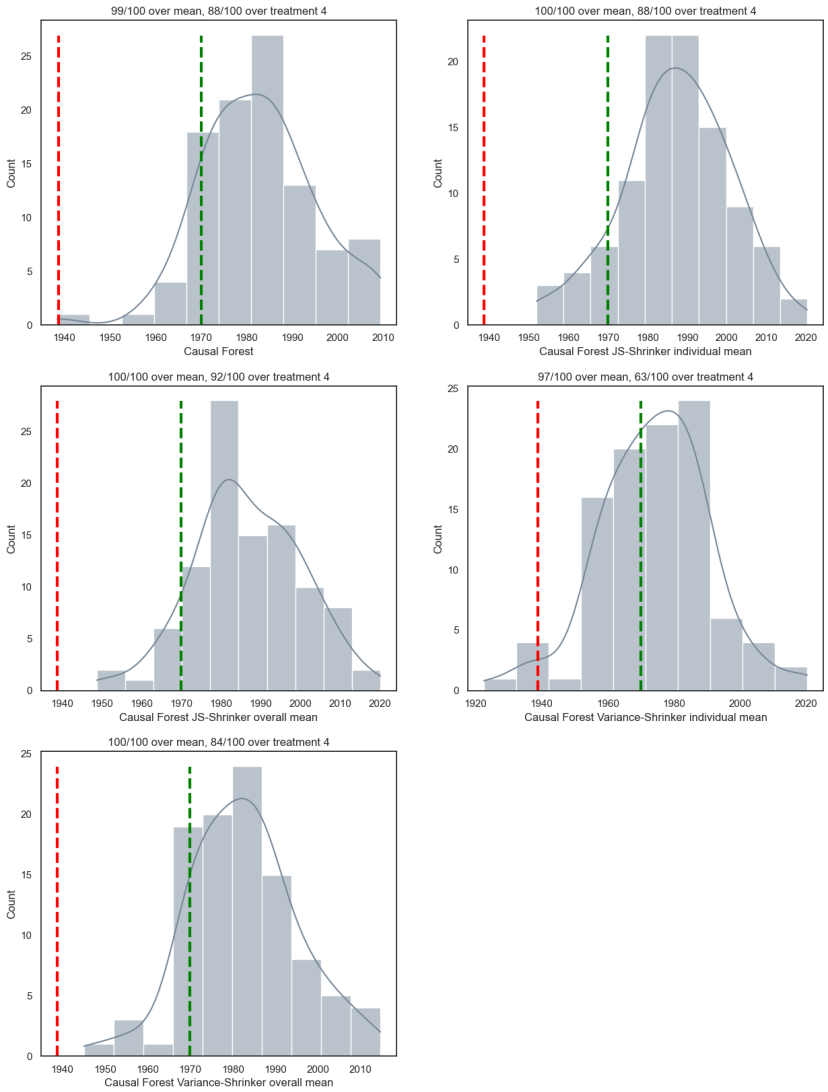

Figure 12: *Hitsch Matching: Using CF, Shrinkage Estimators, and a Subset of Treatments*

The figure shows the distribution of average outcomes from matched observations across 100 individual repetitions of 3-fold cross-validation, using Causal Forests and all four shrinkage methods. Here only the subset of treatments 1, 2, 4 and 5 are considered. The green line represents the average outcome of participants who received treatment four (the loss treatment), equal to 1,970 points. The red line indicates the average outcome of participants who received any of the treatments, equal to 1,939 points. The label "individual mean" refers to shrinkage methods that adjust predictions toward the mean predicted treatment effect of each specific treatment. Conversely, "overall mean" refers to shrinkage methods that adjust predictions toward the average treatment effect of receiving any of the treatments used (as calculated in the training set).

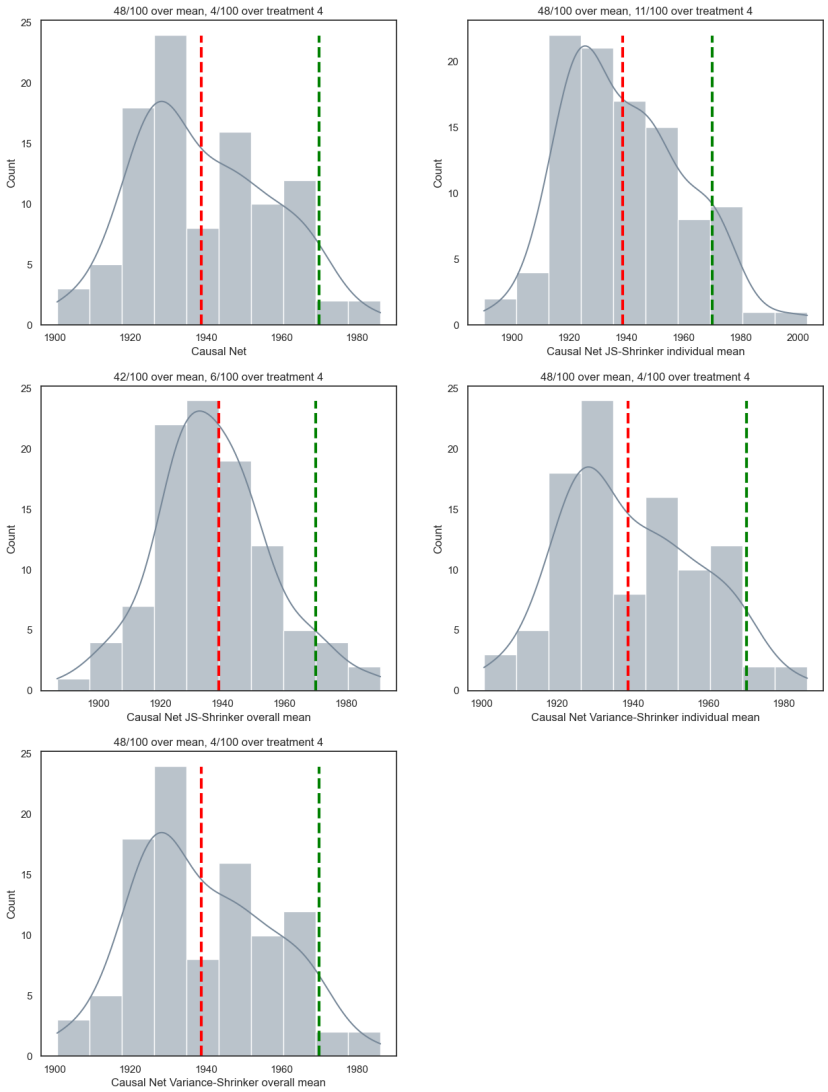

Figure 13: *Hitsch Matching: Using CNN, Shrinkage Estimators, and a Subset of Treatments*

The figure shows the distribution of average outcomes from matched observations across 100 individual repetitions of 3-fold cross-validation, using Causal Neural Networks and all four shrinkage methods. Here only the subset of treatments 1, 2, 4 and 5 are considered. The green line represents the average outcome of participants who received treatment four (the loss treatment), equal to 1,970 points. The red line indicates the average outcome of participants who received any of the treatments, equal to 1,939 points. The label "individual mean" refers to shrinkage methods that adjust predictions toward the mean predicted treatment effect of each specific treatment. Conversely, "overall mean" refers to shrinkage methods that adjust predictions toward the average treatment effect of receiving any of the treatments used (as calculated in the training set).

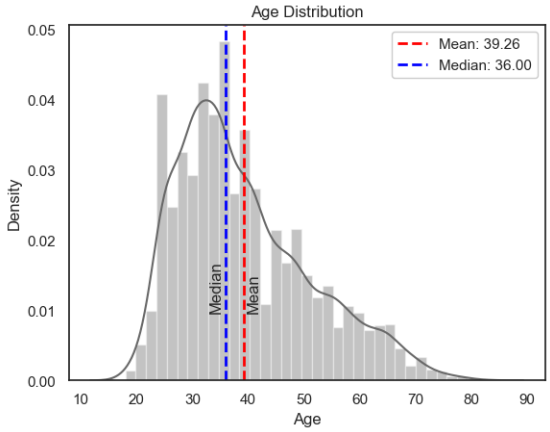

Figure 14: The figure illustrates the age distribution of respondents in the dataset. The histogram reveals a right-skewed pattern, indicating that most respondents are younger, with the frequency gradually decreasing as age increases. The mean age (39.26) slightly exceeds the median age (36.00), suggesting that a smaller proportion of older respondents raises the average age. This distribution suggests that younger individuals are more prevalent in the dataset, with fewer respondents in older age brackets.

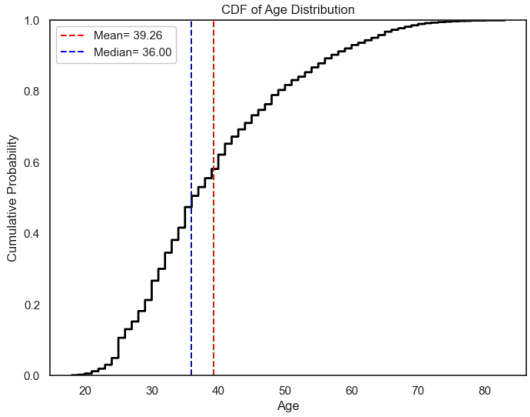

Figure 15: The figure presents the cumulative distribution function (CDF) of respondents' age. The CDF represents the cumulative probability of age, showing the proportion of respondents up to each age. A steeper increase of S-shaped curve around the median shows that a significant portion of respondents are clustered near this age, while the gradual slope afterward highlights the fewer older respondents. The CDF depicts approximately 50% of respondents aged 36 or younger.

9.2 Treatment Details

Outlined below are the treatment titles and the texts the participants were shown with the treatment description.

- **Pay for Performance (PfP)** As a bonus, you will be paid an extra 5 cents for every 100 points that you score.
- **Goal** As a bonus, you will be paid an extra $1 if you score at least 2000 points.
- **Gift & Goal** Thank you for your participation in this study! In appreciation to you performing this task, you will be paid a bonus of $1. In return, we would appreciate if you try to score at least 2,000 points.
- **Loss** As a bonus, you will be paid an extra $1. However, you will lose this bonus (it will not be placed in your account) unless you score at least 2,000 points.
- **Real-Time Feedback** You will receive a bonus that is based on how well you perform relative to others. On your work screen you will see how your current performance compares to that of others who previously performed the task. To that end you will see the percentage of participants who previously performed the task and whom you will outperform at your current speed. You will receive a bonus of $0.02 times the percentage of participants who performed worse than you at the end of the task. That is, you will for instance receive an additional bonus of $1.00 (=$0.02*50) if you perform better than 50% of the participants. The ranking shown on the screen is computed assuming you keep the speed with which you pressed 'a' and 'b' for the past 10 seconds. Your current percentile as well as your currently expected bonus is updated every 10 seconds.
- **Social PfP** As a bonus, you will be paid an extra 3 cents for every 100 points that you score. On top of that, 2 cents will go to Doctors Without Borders for every 100 points.
- **Control** Your score will not affect your payment in any way.

9.3 Model Selection Criteria $\tau - \mathbf{risk}_R$

Using the propensity formula from Equation 4, and assuming the expectation of the outcome given X_i is:

$$m(X_i) = E[Y_i|X_i]$$

, and the (heterogeneous) treatment effect is:

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

Then with $E[\varepsilon_i(W_i)|X_i, T_i] = 0$,

$$Y_i = Y_i^0 + T_i\tau(X_i) + \varepsilon_i \tag{46}$$
$$\varepsilon_i = Y_i - Y_i^0 - T_i\tau(X_i) \tag{47}$$
$$\varepsilon_i - p(X_i)\tau(X_i) = Y_i - [Y_i^0 + p(X_i)\tau(X_i)] - T_i\tau(X_i) \tag{48}$$

as $m(X_i) = E[Y_i/X_i] = Y_i^0 + p(X_i)\tau(X_i)$,

$$\varepsilon_i - p(X_i)\tau(X_i) = Y_i - m(X_i) - T_i\tau(X_i) \tag{49}$$
$$\varepsilon_i = Y_i - m(X_i) - (T_i - p(X_i))\tau(X_i) \tag{50}$$

such that,

$$\varepsilon_i^2 = ((Y_i - m(X_i)) - (T_i - p(X_i))\tau(X_i))^2 \tag{51}$$

Therefore, the minimization of $\tau - \text{risk}_R$ leads to minimization of the squared error term in $Y_i = Y_i^0 + T_i\tau(X_i) + \varepsilon_i$.

9.4 Tuned Hyperparameters

9.4.1 Causal Forest

- *max_features*: 0.2, 0.3,...,0.9, 1.0
- *max_samples*: 0.1, 0.2, 0.3, 0.4, 0.5
- *min_samples_leaf*: 5, 10, 20, 50
- *min_var_fraction_leaf*: 0.1, 0.2, 0.3, 0.4, None
- *max_depth*: 5, 10, 25, 50, 75, 100, None
- *n_estimators*: 1000
- *random_state*: 42

For this model, I use *econml* package in Python.

9.4.2 Causal Neural Network

For the hyperparameter values of causal neural network, I draw the values from the parameters used in the original paper of Farrell et al. (2021).

- *hidden_layer_size / drop_out_rate* - 1 Layer, 60 nodes, 50% dropout rate
  - 1 Layer, 100 nodes, 50% dropout rate
  - 2 Layers, L1: 30 nodes with 50% dropout rate, L2: 20 nodes with no dropout
  - 2 Layers, L1: 30 nodes with 30% dropout rate, L2: 10 nodes with 10% dropout rate
  - 2 Layers, L1: 30 nodes with no dropout, L2: 30 nodes with no dropout
  - 2 Layers, L1: 30 nodes with 50% dropout rate, L2: 30 nodes with no dropout
  - 3 Layers, L1: 100 nodes with 50% dropout rate, L2: 30 nodes with 50% dropout rate, L3: 20 nodes with no dropout
  - 3 Layers, L1: 80 nodes with 50% dropout rate, L2: 30 nodes with 50% dropout rate, L3: 20 nodes with no dropout
- *learning_rate*: 0.1, 0.05, 0.01, 0.001
- *alpha*: 0.01, 0.1, 1 (Regularization Strength parameter)
- *r_par*: 0, 0.3, 0.6 (Mixing ratio of Ridge and Lasso regularization. At 1 equal to Lasso)
- *optimizer*: Adam
- *batch_size*: None
- *max_epochs_without_change*: 60
- *max_nepochs*: 10000
- *seed*: 42

I use Lasso/Ridge kernel regularization, therefore, I standardize the covariates for better learning performance.

List of Figures

List of Tables