Machine Learning and the Propensity Score

Theory and Applications in Development Economics

Mitchell Cameron

2024-08-04

This article evaluates novel approaches to do some really important things.

Table of contents

Preface

This project is submitted in partial fulfillment of the requirements for the Master of Applied Science in Statistics at the University of Otago. My academic background in economics and politics sparked an enduring interest in causal inference, particularly its role in shaping evidence-based policymaking. However, my focus has evolved to include machine learning — a field that, while not traditionally central to economics, offers powerful tools for refining causal analysis.

The motivation for this project stems from my desire to bridge the gap between propensity score methods and machine learning. While the literature is rich with simulation studies, there is a noticeable lack of comprehensive, tutorial-style resources that guide readers through the application of machine learning to propensity score estimation. This project aims to fill that void, providing a practical and accessible exploration of these techniques with approachable theoretical discussion and coded examples in R. As a dedicated user of R, all the packages and methodologies discussed in this project exist in the R ecosystem. Although comparable tools exist in other languages and software.

The intended audience for this work includes individuals with a foundational understanding of causal inference and machine learning, particularly those interested in enhancing propensity score models. However, my discussion extends beyond propensity scores; much of what is covered is relevant to anyone using machine learning for probability prediction.

Acknowledgments

I would like to express my deepest gratitude to several individuals whose support, guidance, and encouragement have been invaluable throughout the course of this thesis.

Presentation Tools

Quarto is an open-source publishing system that enables the creation of dynamic documents, reports, presentations, books, and websites using R, Python, or Julia. It integrates code, markdown, and graphics seamlessly, making it ideal for reproducible research and communication. I the Quarto Book template to create which is hosted on GitHub.

Reading a Quarto Book

Quarto web books have a range of helpful features which are demonstrated below. **Search Feature**

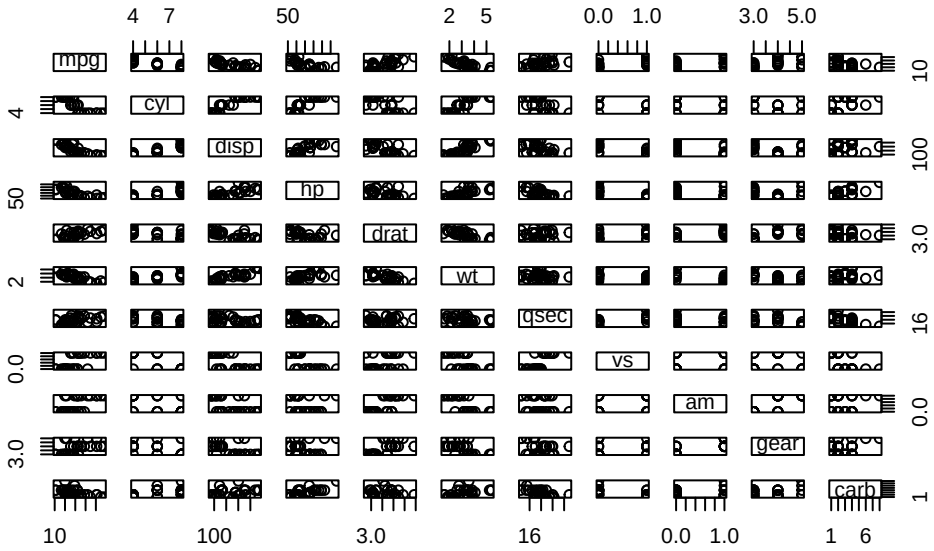
PDF Download

Dark Mode

Folding Code

As code chunks may lead to a clunky presentation with poor readability, code is hidden away and can be down. Additionally, tere is a code show all/hide all option to the right of the table of contents.

```
plot(mtcars)
```



1 Introduction and Background

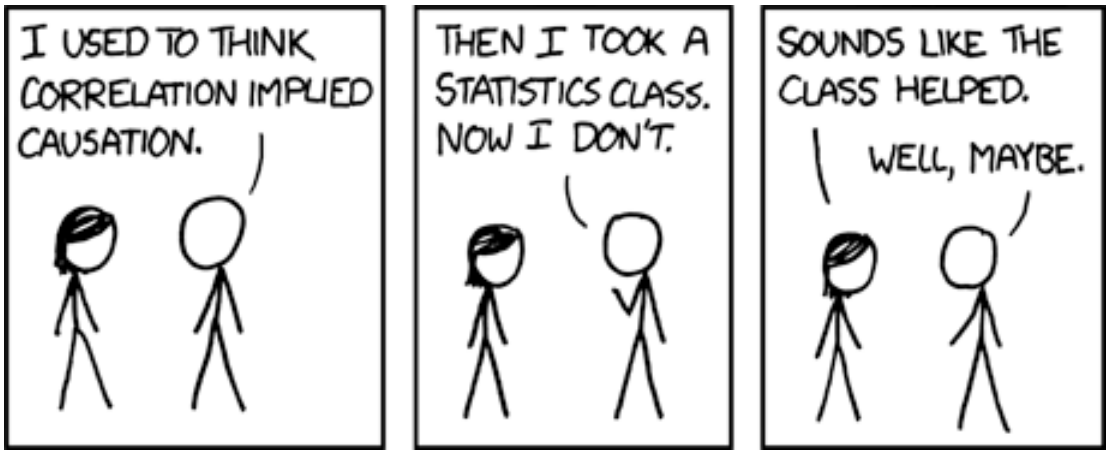


Figure 1.1: The comic plays with the difference between causation vs. correlation. Original: https://xkcd.com/552/

1.1 What is Causal Inference?

Causal inference is a field of study that focuses on identifying and estimating causal relationships between *things*. It goes beyond correlation by establishing a cause-and-effect relationship. Causal inference methods often utilise counterfactual reasoning to estimate the causal effect of an exposure or treatment on an outcome, allowing researchers to make causal claims about the impact of interventions or treatments on outcomes. Such counterfactual reasoning is used unbeknownst every day. For example, if someone misses their bus and thinks, "If I had left home five minutes earlier, I wouldn't have missed it," they are engaging in counterfactual reasoning. In everyday life, policy-making, medicine, and business, understanding the size and nature of a cause of an outcome is essential for decision making and avoiding misleading conclusions. In this background chapter I discuss key ideas in causal inference such as the potential outcomes framework, common estimands, assumptions, graphical displays, and heterogeneous treatment effects.

1.2 Potential Outcomes Framework

The Potential Outcomes Framework, also known as the Rubin Causal Model, was introduced by Rubin (1974) and builds upon the work of Splawa-Neyman (1923). The framework dominates how researchers think about causal inference by formalising counterfactual reasoning. Rubin defines a causal effect as a defined comparison between two states of the world. For each individual, there are two potential outcomes: one if they receive the treatment and one if they do not. The causal effect is the difference between these two potential outcomes. Hence we have two *potential outcomes*, one with the treatment and one without.

The framework is highly flexible and adaptive, extending beyond traditional notions of "treatment" in medical or experimental contexts. It can apply to any kind of intervention, exposure, or condition that could influence an outcome, whether it's a medical treatment, policy change, environmental exposure, or even abstract events like decisions or natural occurrences. Philosophically, the framework aligns with a view of the world that considers reality through alternative scenarios or "what-ifs".

Consider a binary treatment variable, let the treatment for an observation be a random variable, T, with a realisation $t_i \in \{0, 1\}$ under control and treatment. The absence of treatment is refer to as the *control* state. Let $Y_i(1)$ and $Y_i(0)$ be the two potential outcomes for observation i under treatment $T = 1$ and control $T = 0$ respectively. Let the individual treatment effect (ITE), τ_i, be defined as the difference between the two potential outcomes:

$$\tau_i = Y_i(1) - Y_i(0). \tag{1.1}$$

> **Note 1: Fundamental Problem of Causal Inference**
>
> There is a clear problem that only the outcome under either treatment or control is observable. If our observations are on people, then it is logically impossible for an individual to simultaneously both receive and not receive the treatment. For example, if someone takes medication to relieve a headache and their headache improves, it could never know what would have happened if they did not take the medication. This leads to the commonly discussed *fundamental problem of causal inference* - it is impossible to observe both potential outcomes. A counterfactual, the counter to the observed outcome, is infeasible and can never be practicably known.

Let the observed outcome for i be denoted $y_i(1)$ and $y_i(0)$ under treatment or control. Many causal inference methods involve finding or estimating a counterfactual to compare outcomes to sove some variation of Equation 1.1. Let an estimated potential outcomes for i be denoted $\hat{y}_i(1)$ and $\hat{y}_i(0)$ under treatment or control.

1.3 Estimands

In causal inference, there are multiple parameters of interest called and estimand. The preferred estimand depends on the motivating example, discipline, or intended interpretation of a result.

The most basic estimand is the *average treatment effect* or the ATE denoted τ_{ATE} which is the average amount of effect on all individuals in the population regardless of whether they receive the treatment or not. This can be written as:

$$\begin{aligned} \text{ATE} &= E[\tau_{ATE}] \\ &= E[Y(1) - Y(0)] \\ &= E[Y(1)] - E[Y(0)]. \end{aligned} \tag{1.2}$$

Under certain conditions, such as a randomised control trial, Equation 1.2 can be an estimated using:

$$\widehat{\text{ATE}} = \widehat{\tau_{ATE}} = \frac{1}{N_t} \sum_{i=1}^{n} (y_i \mid t_i = 1) - \frac{1}{N_c} \sum_{i=1}^{n} (y_i \mid t_i = 0), \tag{1.3}$$

where N_t and N_c are is the number of treated and control observations.

The second parameter of interest is the *average treatment effect on the treated* or ATT and is the difference (contrast) between the potential outcomes of those who actually receive the treatment. In other words, considering observations where $T = 1$, what is the effect of the treatment? This can be written as:

$$\begin{aligned} \text{ATT} = \tau_{ATT} &= E[\tau \mid T = 1] \\ &= E[Y(1) - Y(0) \mid T = 1] \\ &= E[[Y(1) \mid T = 1] - E[Y(0) \mid T = 1]. \end{aligned} \tag{1.4}$$

The final parameter is the *average treatment effect on the control* or ATC which is similar to the ATT but on those who are actually under control. The ATC is the contrast between the two potential outcomes for individuals which are actually in the control. This is also known as the average treatment effect on the untreated or the ATU. It can be written as:

$$\begin{aligned} \text{ATC} = \tau_{ATC} &= E[\tau \mid T = 0] \\ &= E[Y(1) - Y(0) \mid T = 0] \\ &= E[[Y(1) \mid T = 0] - E[Y(0) \mid T = 0] \end{aligned} \tag{1.5}$$

For the ATT and ATC, no explicity expression exist and estimation is completed using G-methods to obtain contrasts of potential outcomes (see Naimi, Cole, and Kennedy 2017).

1.4 Assumptions in Causal Inference

Assumptions are made for the potential outcomes framework to be logically coherent and for estimands to be identifiable. Firstly, *independence* must be assumed, implying the potential outcomes are independent of T. This assumption is also known as unconfoundedness, ignorability, or selection on observables, and means there is no confounding relationship between the treatment and potential outcomes. This matters as confounding variables can create a spurious relationship between the treatment and the outcome, leading to biased estimates of the treatment's effect. Hence, the treatment assignment should be random, allowing an unbiased estimate. Mathematically independence can be stated as:

$$(Y(1), Y(0)) \perp\!\!\!\perp T. \tag{1.6}$$

Independence implies exchangeability meaning the individuals in the treatment and control groups could be swapped and the potential outcomes are still the same. A weaker assumption is conditional independence that states that assignment into treatment is random conditioned on some X:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X. \tag{1.7}$$

The assumption requires that covariates must be known and measurable which may not always hold. Independence motivates the use of randomisation in experimental contexts as this should guarantee independence. Chapter 2 discusses conditional independence and uses *propensity scores* to "condition" on covariates.

A second assumption is *positivity*. This means that for each i, the probability of being in either the treatment or control group is strictly between 0 and 1. In other words, $\Pr(T = 1 \mid X = x) > 0$ and $\Pr(T = 0 \mid X = x) > 0$. This ensures that all observations have at least some chance of receiving either the treatment or control. If not, it is theoretically impossible to obtain both potential outcomes and so the treatment effect cannot be estimated.

Building on positivity, the fourth assumption is *common support*. This means that the treatment and control groups must overlap in terms of their characteristics. Overlap is crucial because it ensures that for every person in the treatment group, there are similar individuals in the control group—similar in terms of age, gender, income, and other factors. Mathematically, for all of i, if the conditional probability of being treated, $\Pr(T = 1 \mid X = x)$, is near to 1, and $\Pr(T = 0 \mid X = x)$ is near to 0, then there are not compatible cases and there is no *common support*. Without compatible cases, it is not possible to satisfy exchangability and so treatment effect estimates are likely to be biased.

The fifth assumption is *consistency* between the potential outcome and observed outcome. For every i, the observed outcome under treatment equals the potential outcome under treatment. Additionally, the observed outcome under control equals the potential outcome under control.

Mathematically, $y_i(1) = Y_i(1)$ and $y_i(0) = Y_i(0)$ that leads to a switching equation which defines y_i as a function of the potential outcomes:

$$y_i = T_i Y_i(1) + (1 - T_i) Y_i(0). \tag{1.8}$$

Notice the logic of this equation, when $T = 1$ then $y_i = Y_i(1)$ as the second term becomes zero. Similarly, when $T = 0$ then $y_i = Y_i(0)$ as the first term becomes zero.

The final key assumption is called the *stable unit treatment value assumption* or SUTVA. This is a complex way of stating that there is no interference between observations. More specifically, neither potential outcome is affected by the treatment status of any other individual. To borrow terminology from economics, there are no externalities or spillover effects from one observations' treatment status to another observations' potential outcomes.

> **Note 2: Why Assumptions Matter**
>
> In causal inference, especially when working with observational data, it is critical that these assumptions are considered. If these assumptions do not hold, any model, regardless of the modelling assumptions, will not have a causal interpretation. Unfortunately, there are no tests that can confirm if these causal assumptions hold and thus researchers must understand the context and data generating process in which they operate.

2 Propensity Scores and Machine Learning

2.1 A Conventional Approach: Propensity Scores and Balance

In a randomised control trial (RCT), researchers believe treatment and control groups are similar because of randomisation. In this case, the similar groups are compatible and should not have systematic differences. For similar groups, the average treatment effect (ATT) is a contrast of means from Equation 1.3. In observational data, the exposure to a treatment is unlikely to be random, implying there may be systematic differences between groups. Systematic differences refer to consistent variations or disparities between groups in the study. These differences are not due to random chance but rather indicate a pattern or trend, perhaps due to selection-bias. As groups are not comparable, Equation 1.3 leads to a biased estimate of the treatment effect.

For example, consider the causal question: *"How much does obtaining a bachelors degree increase lifetime earnings?"*. Individuals who complete a bachelor's degree are not selected at random for university programs (treatment) and may have different observable attributes than those who do not attend a university (control). Perhaps those who attend university have higher academic abilities, higher motivation, or grew up with parents with higher income. Because of these systematic group covariate differences, a simple comparison of mean income could lead to attributing university attendance as the *cause* of higher incomes when the effect is confounded by the differences in covariates between groups. In this example, the confounding covariates are academic ability, motivation, and parental income that impact the probability of someone obtaining a bachelors degree. This discussion introduces the idea of *covariate balance* which is a key concept behind underlying propensity score methods.

> **i** Note 3: What is Covariate Balance
>
> Covariate balance is the idea that covariates are approximately equivalent across treatment and control groups. If the distribution of each covariate are the same for each group, then those covariates are *balanced*. If covariates are similar across groups, then there should not be any confounding. Equally, similar covariates across groups implies exchangability between groups as the two groups should be similar (thus can be exchanged). There is a conceptual equivalence between covariate balance, unconfoundedness, and exchangeability meaning that Equation 1.6 is satisfied when covariates are balanced.

In bachelor's degree example, suppose that comparable treatment and control individuals are matched together to create balanced pairs. Between these pairs, covariates are balanced such as the same academic ability, motivation, parent income, geographic residence etc. Comparing the balanced matched pairs should result in a robust estimate of a bachelor's degree's impact on earnings because the individuals are exchangeable and satisfy Equation 1.7. The covariates are said to be "conditioned on" by matching individuals on these covariates. However, practically this matching is difficult to perform as exact matches cannot be made as the number of covariates increases. For example, finding two people with the same gender is simple but finding two people with the same gender, age, education, income, motivation, location, experience, and race is nearly impossible. Thus, there is a *dimensionality* problem as the dimension of the number of covariates increases.

Rosenbaum and Rubin (1983) offer a valuable tool for analysing observational data called the propensity score. The propensity score is the probability of treatment assignment conditioned on observed covariates. Essentially, the propensity score reduces the dimension of the number of covariates to a single dimension to avoid the dimensionality problem. Let the propensity score be denoted as $e(X)$ and be expressed as:

$$e(X) = P(T = 1|X). \tag{2.1}$$

A prediction of the probability of treatment based on the covariates is the best summary of each covariate. The covariate imbalance between bachelors degrees and controls arose from people self-selecting themselves into a bachelors degree programme because of these covariates. For example, people with higher motivation and academic ability are more likely to go to university. If it is the covariates that impact the probability of going to university, then a prediction of the probability of going to university based on these covariates should summarise the covariate effects.

Conditioning on this propensity score should balance the data and meet the conditional independence assumption stated in Equation 1.7. There are many sources that offer a comprehensive guide to propensity score methods such as (Cunningham 2021, chap. 4) who provides applications and coded examples in R, Python, and Stata.

> **Note 4: Balance and Propensity Scores**
>
> Note that an RCT will satisfy Equation 1.6 as randomisation implies the potential outcomes are independent of the treatment assignment. Propensity score methods aim to satisfy Equation 1.7 as the potential outcomes are independent of the treatment status conditioned on some covariates.
> The propensity score is conditioned on the covariates because the covariates are the predictors of the propensity score. Conditioning on the propensity score aims to replicate an RCT in observational data by balancing covariates between groups. When observations are conditioned on their propensity score, differences in outcomes can be confidently

attributed to the treatment itself, rather than to pre-existing differences in covariates.

Two common methods that use propensity scores are propensity score matching (PSM) and inverse propensity weighting (IPW). PSM creates matched sets with similar propensity scores. IPW creates a balanced pseudo-population, where observations are weighted on the inverse of the propensity score. The pseudo-population is created by up-weighting observations with a low propensity score and down-weighting observations with a high propensity score.

At a high level, the conditioned property of the propensity score is translated into a model by using PSM or IPW. King and Nielsen (2019) provide a notable criticism of propensity score matching, which is a very interesting read. In the following examples, IPW is used due to theoretical advantages and ease of software implementation.

2.1.1 Assessing balance

Balance assessment is an important step to ensure that conditioning on the propensity score has been successful. A commonly recommended measure of covariate balance is the standardized mean difference or SMD. This is the difference in the mean of each covariate between treatment groups standardized by a standardization factor so that it is on the same scale for all covariates.

SMDs close to zero indicate good balance. P. Austin (2011) notes that 0.1 is a common "threshold" for determining if a variable is balanced. This threshold is a guideline to the approximate region that indicates a covariate is balanced and should not be interpreted as a binary rule.

2.1.2 Propensity Score Modelling with Logistic Regression

A conventional propensity score model uses logistic regression to predict a probability between 0 and 1. Models may be specified to include interaction terms and polynomial terms so the model captures complex trends in the data. There are a range of approaches for specifying a propensity score model, but the process is a heuristically driven art rather than science. (Brookhart et al. 2006; Heinrich 2010). One suggestion is to include two-way interaction terms between covariates and squared terms and then remove terms which are not statistically significant. Notably, many researchers do not discuss the specification of propensity models in their papers. P. C. Austin (2008) reviews 47 papers that use propensity scores and find few perform adequate model selection, assess balance, or apply correct statistical tests.

It is important to note that the true propensity score is never observable. A propensity score that is close to the theoretically true probability is well calibrated. Poorly calibrated propensity scores may result in poor balance and biased estimation of the treatment effect. Propensity scores may be poorly calibrated as covariates may be omitted by error, poorly measured, or

be unobservable. Logistic regression may not predict calibrated scores if the true relationship is non-linear or involves complex interactions between covariates. Another important note is that the propensity model itself does not have an informative *causal* interpretation. In logistic regression, the coefficients are the log-odds of the treatment assignment for each variable which is not informative of the desired estimand.

The first application of machine learning in causal inference was to predict propensity scores. Despite this, logistic regression still appears to be the most common model for predicting propensity scores.

2.2 Probability Machines: Probability Theory and Machine Learning

Supervised machine learning usually focuses on classifying observations into groups, or predicting continuous outcomes. Probability prediction is a hybrid of these tasks, aiming to predict the continuous probability an observation belongs to a certain class.Machine learning methods that predict probabilities are sometimes called probability machines.

Probability machines are valuable in applications requiring calibrated probability predictions. For example, probability machines can predict loan defaults or other adverse events in finance. In marketing, they estimate the likelihood of customer response to a campaign. Gamblers and bettors want robust probability predictions to enhance their betting strategies. Probability machines can be applied wherever calibrated probability predictions are needed.

Probability machines offer many advantages over parametric methods like logistic regression:

1. **Improved Calibration**: Probability machines often provide better-calibrated predictions by capturing complex data relationships.
2. **Flexible Modelling**: Unlike parametric methods like logistic regression, probability machines don't rely on assumptions of additivity or linearity, allowing them to model intricate relationships that parametric models miss.
3. **Efficient Feature Selection**: These machines automatically select features, making them ideal for high-dimensional datasets where manual selection is impractical.
4. **Handling Missing Data**: Probability machines handle missing data robustly, minimizing the need for extensive data reprocessing and imputation.
5. **Simplified Data Exploration**: By exploring complex data structures in a data-driven way, probability machines simplify model specification. For instance, tree-based models remain unaffected by adding squared or interaction terms, streamlining the modeling process.

In causal inference, probability machines can predict better calibrated propensity scores and better estimate treatment effects. This discussion aims to clarify the use of probability machines in causal inference given the unique requirements of propensity score specification. Probability machines are theoretically complex and there are unanswered questions in this space.

> **i** A Particularly Important Method: Classification and Regression Trees
>
> Moving forward, a particularly important model is the Classification and Regression Trees. L. Breiman et al. (1984) introduces method, commonly known as CART, that partitions data according to a splitting criterion, resulting in an "if this, then that" interpretation. CART models are also widely known as a decision trees. The splits are recursive, meaning splits are applied upon previous splits, such as trees breaking into branches and then leaves. The splits are also *greedy* as each potential split only considers information available at that split instead of past or future splits. Each parent node is split to create two child nodes and the final nodes of a CART model are called terminal nodes.
> For example, when classifying pets into cats versus dogs, the first split might be *"if barks"* and the second is *"heavier than 5 kg"*. The tree says *If it barks and is heavier than 5 kg, then it is a dog.*
> A single classification tree typically uses the Gini index to determine its splits. Each split aims to maximize node purity, meaning the nodes contain the highest possible proportion of one class. The Gini index measures impurity, with lower gini values indicating higher purity. Intuitively, the aim of a classification tree's loss function is to minimize the misclassification rate of observations. By selecting splits that reduce the Gini index, the tree minimise classification errors and accuracy.

2.2.1 Choice of Loss Function and Probability Prediction

The loss function measures the difference between a model's predictions and the actual target values, serving as a measure of a model's performance. The best model exists at the minimum of the loss function. Different loss functions influence a model's behaviour so the choice of loss function is important. Classification models predict the category that each observation belongs to not the probability of each class. For instance, in fraud detection, banks use classifiers to distinguish between fraudulent and routine transactions. Many classification loss functions minimize classification errors and improve accuracy as this results in the best classification. A loss function like the Gini index is effective for classification problems but it is unclear if this applies to probability problems. In other words, minimizing misclassification error may not lead to accurate probability predictions.

At a high level, to classify an observation, x_i as an A or B, a model needs to determine if $\Pr(x_i = A)$ is less than or greater than 0.5. If $\widehat{\Pr}(x_i = A) > 0.5$, then it is more likely to be an A and if $\widehat{\Pr}(x_i = A) < 0.5$ then it is more likely to be a B. Thus, if x_i is an A it is trivial if $\widehat{\Pr}(x_i =$

A) is 0.51 or 0.99 as this makes no difference to the classification as an A. But the difference between $\widehat{\Pr}(x_i = A) = 0.51$ and 0.99 is extreme for a probability machine. It is important to understand that classification models are optimized for classification accuracy rather than probability prediction. This distinction affects the performance of ensemble methods like random forests or bagging ensembles that use classification trees for probability prediction.

2.2.2 Bagging and Random Forest as Probability Machines

> **Note 5: A Quick Note on Ensemble Learning**
>
> Ensemble learning refers to a general framework of machine learning that combines together multiple simple models to create a better model. The philosophy behind ensemble learning is rooted in the wisdom of crowds principle, where the collective decision of multiple models often outperforms that of individual models. Often, ensemble learning methods use multiple CART models.
> Bagging ensembles, random forest, and gradient boosting are all machine learning methods that combine many weaker models and so are all examples of ensemble learning.

Across an entire bagging ensemble (see Leo Breiman 1996) or random forest (see Brieman 2001), class probabilities are determined through a *vote count* method. Within that ensemble, each tree makes a class prediction based on the majority class in a terminal node. For instance, if x_i lies in a terminal node where 80% of the observations are classified as an A, that *individual tree* will classify x_i as A. The ensemble's overall prediction for x_i is derived from the proportion of trees that classify x_i as A or B. Thus the ensemble *counts votes* for each class across the ensemble. Let T be the total number of trees and b_t be the t-th tree in the ensemble. Let $\mathbb{I}(b_t(x_i) = A)$ be the indicator function that returns 1 when b_t predicts that observation x_i belongs to class A. The probability of class A for observation x_i is calculated as:

$$\Pr(x_i = A) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}(b_t(x_i) = A). \tag{2.2}$$

Olson and Wyner (2018) notes a potential bias towards predictions of 0 or 1 when trees in a bagged ensemble or random forest are highly correlated. Using a vote count method, an ensemble will bias predicted probabilities towards $\hat{P}(x_i = A) \in \{0, 1\}$ when trees are highly correlated. Imagine that each tree in the ensemble is perfectly correlated implying that each tree will make the the same class prediction for each x_i. For such an ensemble, the predicted probabilities will will be exactly 0 or 1 using a vote count. Of course an ensemble of identical trees is unrealistic but the intuition still applies in the real world where ensembles may have some degree of correlation. The larger the correlation, the more the probability predictions will exhibit a *divergence bias* towards 0 and 1. Notably, divergence bias is not problematic in

classification applications, as a larger number of trees correctly classifying the observation is encouraging.

If x_i has a known membership of A, and an unknown $\Pr_{\text{true}}(x_i = A) = 0.6$, the ensemble might classify x_i correctly 90% of the time leading to $\widehat{\Pr}(x_i = A) = 0.9$. As a probability machine, the ensemble has overestimated the probability by 0.3 even though a 90% classification accuracy is excellent. To predict $P_{\text{true}}(x_i = A) = 0.6$, an ensemble needs to incorrectly classify x_i in 40% of its trees. However, bagging ensembles and random forests are designed to maximize classification accuracy and there is no incentive for the model to intentionally achieve a specific misclassification rate that aligns with the true probability.

To reduce tree correlation, bagging ensembles use bootstrap aggregation and train each tree on a randomly selected bootstrapped sample of the data. Random forests further reduce tree correlation by considering a random number of variables at each split, commonly referred to as *mtry* in software implementations. When *mtry* is equal to to number of predictors, the model considers all variables at each split and the random forest is equal to a bagging ensemble. A lower *mtry* should reduce the correlation between trees and decrease divergence bias as the structure of the tree is modified by the selected variables at each split. However, a lower *mtry* also introduces other theoretical problems.

Consider the scenario where the binary outcome (treatment and control) of the ensemble is strongly related to a single predictor and weakly related to other noisy predictors. If *mtry* is low then each split may not consider the strong predictor and more commonly splits on weak or noisy predictors. Each predictor has a chance of $\frac{mtry}{\text{number of predictors}}$ of selection at each split implying a lower *mtry* decreases the chance of a splitting on the strong predictor. Splits on the weak or noisy predictors may not result in a meaningful increase in node purity and successive splits may result in impure terminal nodes that poorly predict the class of x_i in each tree. Such an ensemble may have highly unstable probability predictions.

Additionally, consider there is a class imbalance and the majority of observations are classified as A not B. The terminal nodes of each tree within an ensemble are more likely to contain the majority class. Consequently, there is a *majority class bias* as each tree in the ensemble is more likely to misclassifying an observation as an A because the terminal nodes have a higher proportion of A due to the higher proportion of A's in the data overall.

i Note 6: Class Imbalance and Machine Learning

When there is a difference in the number of observations in each class, this is called class imbalance. It is important to note that majority class bias exists in conventional classification tasks with machine learning. Bagging ensembles and random forest are well known to be sensitive to class imbalance meaning that class predictions are biased towards the majority (see Bader-El-Den, Teitei, and Perry 2019).
However, the class imbalance problem is particularly notable when predicting probabilities. The probability prediction from a vote count method is very sensitive to a change

in the votes from each tree. Suppose that balanced data results in 80/100 trees classifying B as B and imbalanced data (more A than B) reduces correct classifications of B to 60/100. This results in a 20% margin of error in probability estimates but the classification remains as B.

Individually, a low *mtry* can lead to unstable probability predictions and class imbalance can create bias towards the majority class. But probability machines are particularly effected when there is both a low *mtry and* class imbalance. Because successive noisy splits (relating to a low *mtry*) result in impure child nodes, the effects of majority class bias are exaggerated. Without the ability to separate the classes, the majority class will dominate terminal nodes. If the ensemble was able to split on informative covariates each time (*mtry* is higher), then it should still be able to create pure splits even when there is some class imbalance. In other words, if there is a small class imbalance, reducing *mtry* may reveal majority class bias not visible at higher *mtry*'s. Equally, if there is low *mtry*, then even a small class imbalance can lead to majority class bias.

To exemplify these theoretical points, the National Supported Work (NSW) programme is a commonly discussed dataset in causal inference. The data results from a randomized controlled trial with 445 total participants, 185 in the program group, and 260 in the control group, so the true probability of treatment for each individual can be calculated as $185/445 = 0.42$ or 42%. Thus, the "best" calibrated probability estimates will be close too 42% for all observations. Further information about this data is found in Section 6.1.1.

Randomisation should ensure that the probability of treatment is independent of the predictors and so all predictors should be noisy or weak. Although Figure 2.2 and Table 3.1 suggest some covariates do have a greater impact on the probability of participating in the programme, which echoes research by Smith and Todd (2005) who suggests that self-selection bias is evident in the NSW data.

Figure 2.1 shows both divergence bias and majority class bias using `randomForest()` to fit both the random forest and bagging ensemble. Recall that a bagging ensemble is a random forest model when *mtry* is equal to the number of predictors and so specifying `mtry = 7` in the `randomForest()` function fits a bagging ensemble. Additionally, logistic regression using the `gbm()` function provides a meaningful comparison.

Figure 2.1 shows the logistic regression model has identified a central tendency and most propensities are between 0.25 and 0.75 which roughly aligns with the true probability. The bagging ensemble has clear evidence of divergence and the majority of predictions are outside 0.25 and 0.75 which is likely related to tree correlation. For the random forest with $mtry = 1$, a significant number of the treatment and control observations are centred near the control area ($T = 0$) with a wide range of other predictions. Recall that the control group is the majority class. Reducing *mtry* from 7 to 1 reveals the majority class bias reinforcing the theoretical discussion that a combination of low *mtry* and class imbalance is especially troubling. The models over predicts the majority class and has unstable predictions otherwise. Both random

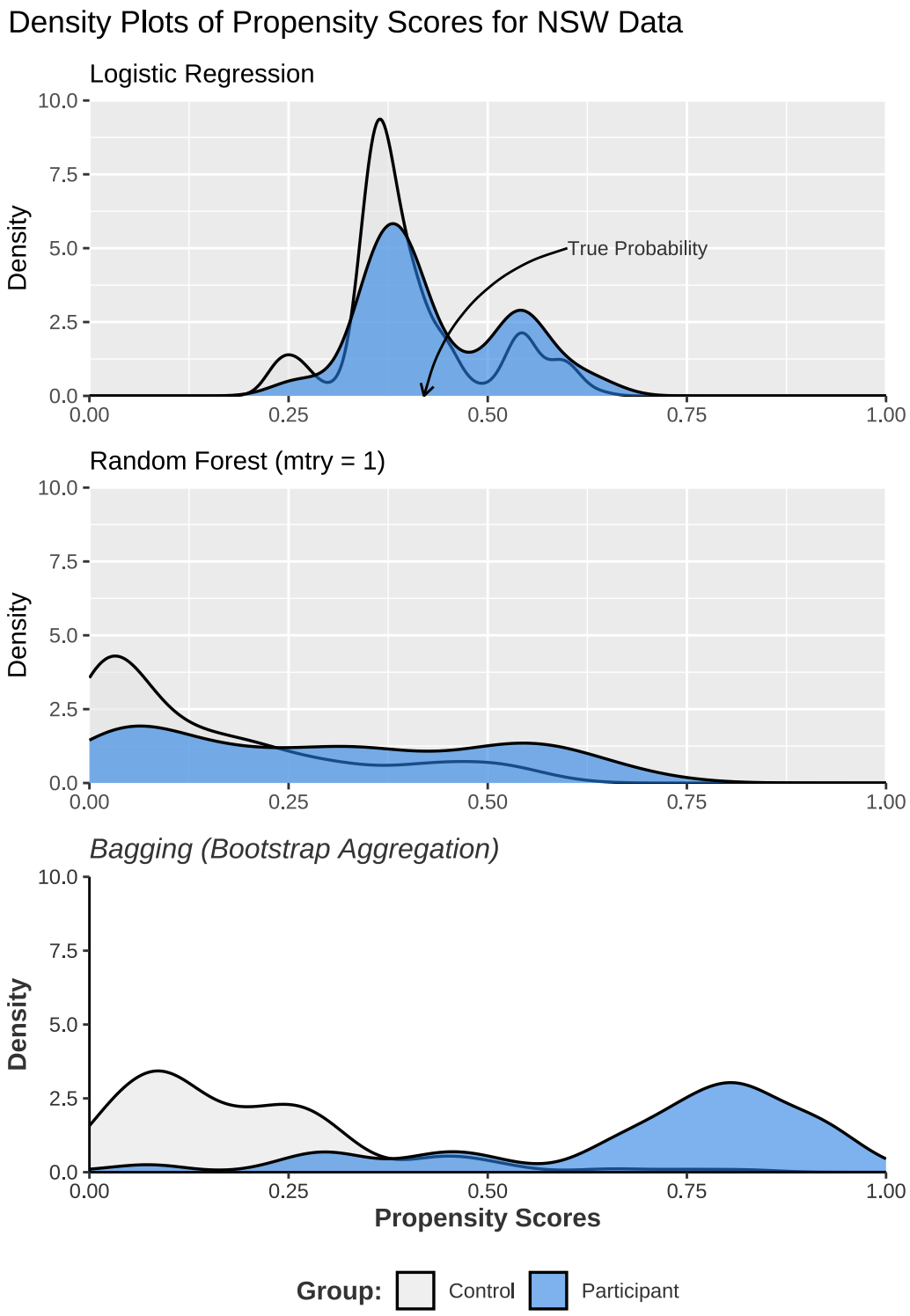


Figure 2.1: Compares the density estimates of the propensity scores for control and participant groups in the National Supported Work programme. `randomForest()` fits a random forest with `mtry = 1` and bagging ensemble with `mtry = 7`. The default values of `ntree = 500` and `nodesize = 1` are used. A logistic regression model is included for a comparison.

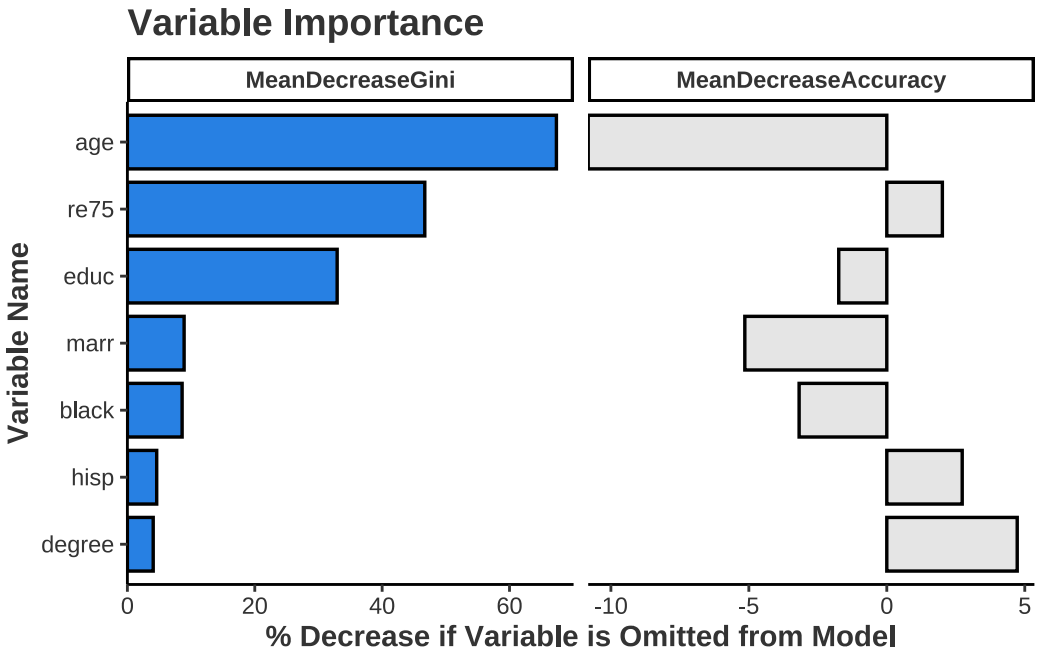


Figure 2.2: Compares the variable importance assigned to each variable from a bagging ensemble fitted on data from the National Supported Work programme. `randomForest()` fits a baggin ensemble with `mtry = 7` with default `ntree = 500` and `nodesize = 1`.

forest and bagging ensembles have performed poorly compared to the true probability of 0.42%.

The tuning of *mtry* faces double jeopardy and is another important area of discussion in probability machines. The selection of *mtry* is typically carried out in with a classification loss function such as accuracy or out-of-bag error. Olson and Wyner (2018) compares tuning *mtry* measured by classification accuracy and mean square error of known simulation probabilities and finds that the optimal value of *mtry* for classification differs from probability prediction.[1] In other words, if a grid search finds that $mtry = 3$ is optimal for a classification task, this does not imply that $mtry = 3$ is optimal for predicting probabilities so the tuning of *mtry* is difficult.

Random forests and bagging ensembles seem to be troubled as probability machines but this does not mean that bagging and random forest cannot perform well. In various simulation studies, they perform excellently as discussed in Section 2.3. Perhaps the nature of the data is informative for the potential success of a random forest or bagging ensemble.

Heuristically, divergence bias and majority class bias will most affect a probability machine when there is considerable overlap of true probabilities between groups. Recall the meaning of *common support* and *overlap* from Section 1.4. If there is overlap and a central region of true probabilities, then the effects of divergence bias may be very pronounced. Similarly, common support may make it even harder to increase purity in child nodes, as the covariates will lack clear split points. When combined with weak predictors relating to a low *mtry*, the terminal nodes of each tree may be relatively impure leading to a majority class bias. Alternatively, if true probabilities exist near 0 or 1 and there is a clear separation of class, divergence bias may trivially effect probability estimation as the probabilities already exist in that region. If there is a clear separation of class, then weak predictors relating to a low *mtry* may still create meaningful splits and pure terminal nodes. It is worth noting that propensity score methods require datasets with overlap to meet the assumptions required to determine causality.

2.2.3 Gradient Boosting Machines as Probability Machines

Moving beyond classification trees in random forests or bagging ensembles, Friedman (2001) introduced the *Gradient Boosting Machine* (GBM). A GBM sequentially constructs CART trees to correct errors made by previous trees. Employing a gradient descent process, each new tree is fit on the pseudo-residuals of the previous iteration. This means that with each iteration, the GBM takes a gradient step down the global loss function, incrementally minimizing the loss function to reach a minimum. The update rule for the model after each iteration can be expressed as:

$$\hat{p}_i^{(t)} = \hat{p}_i^{(t-1)} + \lambda \cdot b_t(x_i), \tag{2.3}$$

[1]Note that tuning *mtry* for the mean square of probability prediction is only possible by design of the simulation study and is not possible in applications, as the true probability is unknown.

where λ is the learning rate, and $b_t(x_i)$ is the b-th regression tree fitted on the pseudo-residuals of the previous regression tree. In words, the current overall iteration t, is a combination of the previous model plus the current iteration scaled by a learning rate.

A learning rate controls the contribution of each weak learner to the final model. By using a small learning rate, the machine learns slowly so that it can slowly descend the loss function. This allows for finer adjustments during the iterative process to better capture patterns in the data.

GBMs can be generalized to many different applications by minimizing a different loss function which can be specified as any continuously differentiable function. For binary outcomes, a GBM employs multiple regression trees and a logistic function to transform regression predictions into probabilities. Specifically, the logistic function used is:

$$\hat{p}_i = \frac{1}{1 + \exp(-\text{model}(x_i))}. \tag{2.4}$$

This logistic function is the same as in logistic regression, so a GBM with a binary class is sometimes called boosted logistic regression. The ensemble aims to minimize the Bernoulli deviance, which is equivalent to maximizing the Bernoulli log-likelihood with logistic regression. Maximizing the log-likelihood ensures that the predicted probability distribution is as close as possible to the true probability distribution given the data. The full GBM model, $f_T(x)$ after T iterations can be written as:

$$f_T(x_i) = b_1(x_i) + \lambda \sum_{t=1}^{T} b_t(x_i). \tag{2.5}$$

Inside a base tree, each split considers all variables and makes the most informative split to descend the loss function using gradient descent. GBMs utilize many weak learners, such as a regression tree with a single split called a regression stump. However, additional splits enable the model to capture interactions between terms, which may increase probability calibration in complex or high-dimensional datasets.

By outputting probability predictions and avoiding the flaws of vote methods in other ensemble techniques, as well as allowing a probability distribution-based loss function optimal for probability prediction, GBMs stand out as a highly effective probability machine. Since GBMs predict probabilities from a logistic function, they avoid problems associated with a vote count method. Additionally, there are no difficult parameters to tune, such as *mtry* in a random forest. The implementation and workflow to fit a GBM for propensity scores is discussed in Section 3.1.

Figure 2.3 shows the propensity scores resulting from a GBM model using the `gbm, package` on the NSW data provides. `gbm` is a notable performance improvement to random forest and

bagging shown in Figure 2.1. Recall that a "better" model would predict probabilities near to 42% as this is the treatment/control share in the randomised NSW data.

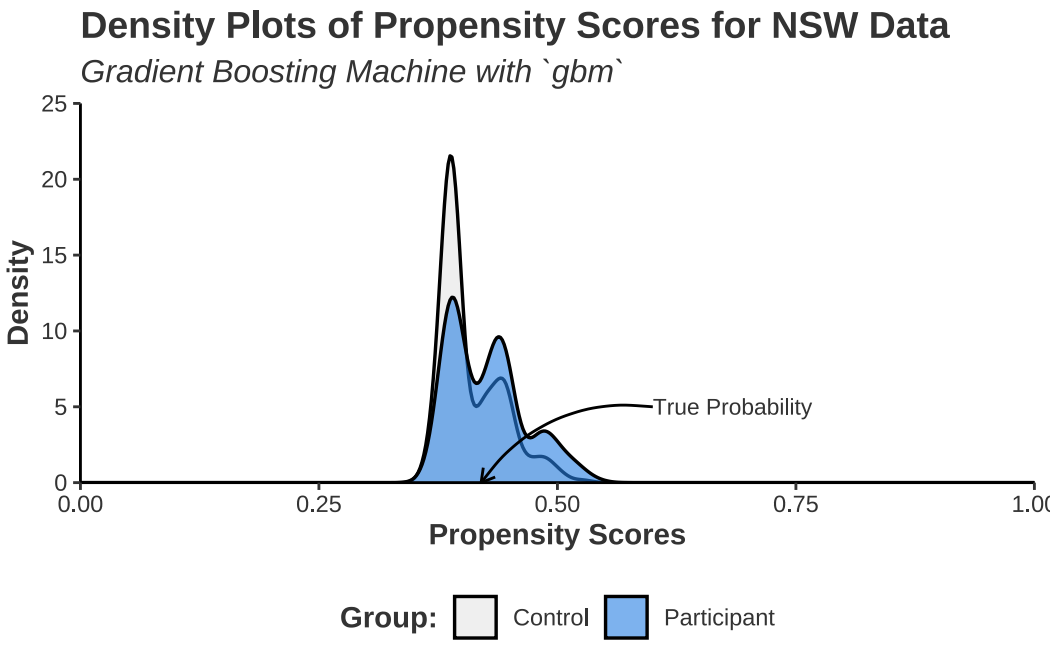


Figure 2.3: Density estimates of the propensity scores for control and participant groups in the National Supported Work programme using the `gbm` package with `distribution = "bernoulli"`, `data = nsw_data`, `n.trees = 10000`, and `shrinkage = 0.0001`.

2.2.4 Overfitting

Overfitting is a common concern when fitting machine learning models, as models can capture noise and random variations in the training data. An overfit model will typically show excellent performance on the training data but will perform poorly on new, unseen data because it cannot generalise beyond the specific patterns of the training set. For instance, consider a machine learning algorithm used by a bank for fraud detection. In this scenario, an overfit model would struggle to classify transactions correctly as it has learned the noise and specific variation in the training data rather than the underlying patterns of fraud. Cross validation or test/train splitting can prevent overfitting to ensure a model can generalise to unseen data.

However, the model is not required to generalise a propensity score model as different datasets will have a different model. Instead, the emphasis of predicting propensity scores is to create balance in the data. A model is effective if it balances covariates between groups, even if it is overfit in a conventional sense.

> **Note 7: Overfitting in Logistic Regression**
>
> There is limited research on how overfitting a logistic regression model affects estimating treatment effects. In logistic regression, overfitting occurs when there are too many parameters and so the maximisation of the log-likelihood function is difficult because of noise. One study that investigates overfitting in this context is Schuster, Lowe, and Platt (2016), who suggest a general rule that the number of observations per parameter should be between 10 and 20. When overfitting occurs, the variance of the estimated treatment effect increases because noise amplifies the magnitude of the coefficients, resulting in a small bias towards 0 or 1 because of properties of the logit function. Specifically, when using (non-augmented) propensity score weighting, the estimate of the treatment effect will have high variance as propensity scores close to 0 or 1 receive artificially inflated weighting.

Lee, Lessler, and Stuart (2010) simulates a comparison of machine learning methods for propensity score prediction and finds that an overfit CART model performs better than a pruned CART model in terms of balance and treatment effect estimation bias. While not conclusive, this suggests that conventionally overfit trees are appropriate and potentially beneficial for propensity score modelling.

If overfitting was to occur, this could be interpreted as balance between groups getting worse decreases with a higher model complexity. Although various software packages use a stopping rule to prevent this. As conventional advice states, creating balance should be the aim of estimating propensity scores.

2.3 Comparison of Machine Learning Algorithms: Simulation Results

A small body of simulation studies benchmarks probability machines for predicting propensity scores (see McCaffrey, Ridgeway, and Morral 2004; Setoguchi et al. 2008; Lee, Lessler, and Stuart 2010; Cannas and Arpino 2019; Tu 2019; Goller et al. 2020; Ferri-García and Del Mar Rueda 2020). Although these studies tackle the same problem, differences in simulation design and model implementation lead to a diverse range of perspectives on this issue. This variety reflects the complexity of the propensity score prediction.

Tu (2019) compares logistic regression, boosting, bagging, and random forests across different sample sizes, conditions of linearity and additivity, and treatment effect strengths. Boosting achieves the lowest bias ATE estimate in most scenarios and the lowest mean square error in all scenarios. Bagging ensembles and random forests perform poorly in both ATE estimate bias and MSE. The author notes that poor performance in bagging ensembles is likely due

to correlated trees in the ensemble, leading to divergence bias. Random forests perform significantly better than bagging but both methods performed worse than boosting or logistic regression.

Despite poor theoretical properties as a probability machine, Lee, Lessler, and Stuart (2010) find that bagging results in the lowest standard error across many datasets.[2] This result is not surprising given that the bagging ensembles are trained on bootstrapped datasets, leading to lower variance and standard error. Although, this advantage is not likely of practical interest given that the small performance gain in standard error is at the expense of a considerable increase of bias.

Additionally, Lee, Lessler, and Stuart (2010) finds that logistic regression performs well in simple data structures with comparable bias to boosting and random forest, but with larger standard errors. In complex data structures, boosting shows low bias and outperforms logistic regression while maintaining low standard errors. Consequently, the study concludes that boosted CART achieves the best 95% coverage in all simulation scenarios, with 98.6% coverage.[3]

Cannas and Arpino (2019) also undergo a simulation study to assess machine learning methods for propensity score prediction. They compare logistic regression, CART, bagging ensembles, random forest, boosting, neural networks, and naive bayes and find that random forest, neural networks, and logistic regression perform the best. Notably, the simulation design only performs hyperparameter tuning for CART, random forest, and neural networks but not either of their boosting implementation. [4] This is a weakness of their study design and thus their findings may be more informative of the relative performance of tuned versus untuned models. Although, the finding that random forest performs well when tuned is significant.

Goller et al. (2020) adds diversity to the simulation study literature by exploring an economics context, experimenting with imbalances between treated and control observations, and incorporating LASSO and probit models.[5] [6] Probit regression achieves the best covariate balance, with LASSO also performing well. In contrast, the random forest model performs poorly,

[2] In this case, the standard error is the dispersion of the standardised mean difference (effect size) across 1000 simulated datasets.

[3] In this context, the coverage is the proportion of times that the true treatment effect is within the 95% confidence interval across the number of simulations. This author implements 1000 simulations of each scenario.

[4] Cannas and Arpino (2019) provide a replication package for their simulation study online and their hyperparameter tuning is process transparent. The authors fit two GBMs using the `twang` and `gbm` package in R. The hyperparameter values provided to these untuned boosting models are contrary to heuristics and may lead boosting to perform poorly regardless of theoretical benefits discussed in Section 2.2.3.

[5] Goller et al. (2020) calculates the bias of the treatment effect using the average of the estimates from logistic regression, random forest, and LASSO models as the *true* treatment effect. Thus, the covariate balance table offers a clearer view of each method's performance.

[6] Tibshirani (1996) introduces LASSO regularization, short for Least Absolute Shrinkage and Selection Operator, is a technique used in linear regression to prevent overfitting by penalising the absolute size of the coefficients. It adds a penalty term to the ordinary least squares objective function, meaning that some coefficients may "shrink" to zero.

showing imbalance statistics with several orders of magnitude higher than those of probit or LASSO. To perform feature selection, a probit model with many interactions and polynomial terms is specified, and a LASSO penalty shrinks covariate coefficients to zero. Probit regression stands out for its superior covariate balance, while LASSO also delivers satisfactory results. The random forest model underperforms with significantly higher imbalance statistics compared to probit and LASSO.

Based on a review of the literature, the findings can be distilled into five important points:

1. Probability machines can predict propensity scores with excellent performance and their implementation should be considered in most scenarios. Although, a logistic regression approach may be preferred because of simplicity while still providing adequate performance in simple data structures.
2. In cases of non-linearity or non-additivity in the data, probability machines often achieve better covariate balance and lower bias of treatment effect estimates than logistic regression. This is significant as propensity scores are frequently used in observational studies with complex data structures (Rosenbaum and Rubin 1983).
3. Bagging ensembles perform poorly, a finding replicated across multiple studies.
4. Random forests can perform excellently when hyperparameters are satisfactorily tuned.
5. Further research should consider parametric methods with LASSO, Ridge, or Elastic Net penalties to assist in feature selection. Simulation study evidence for predicting propensity scores is limited despite attractive properties of these methods.
6. A tuned GBM stands out with strong theoretical support, excellent simulation performance, and superior software implementation and documentation. Specifically, this GBM will use the Bernoulli deviance as a loss function due to theoretical benefits. Implementations of GBMs such as AdaBoost.M1 have no simulation study evidence.
7. A good practical approach seems to be a trial-and-error approach of fitting multiple model specifications, then considering covariate balance for each model.

Short concluding remark

3 Tutorial: Implimentation, Workflow, and Example with `WeightIt` and`gbm` in R

Based on Friedman (2001), the `gbm` package implements a *Generalized Boosting Machine*. Here, the "generalized" is because the package provides generalisations of the boosting framework to other distributions such as Bernoulli, Poisson, and Cox-proportional hazards partial likelihood of class probability predictions. Although this implementation very closely follows Friedman (2001) who introduced the gradient boosting machine. `gbm` also supports stochastic gradient boosting, which performs random bootstrap sampling for each tree using the `bag.fraction` parameter.

To fit and tune a GBM for propensity scores, wrapper packages facilitate optimal hyperparameter tuning for covariate balance. An effective approach involves fitting the model and computing balance statistics at each hyperparameter combination. Since the `gbm` package does not support this type of tuning, a wrapper package like `WeightIt` is necessary. `WeightIt` allows for hyperparameter tuning based on covariate balance and inverse propensity weighting (IPW). `WeightIt` supports hyperparameter turning of `shrinkage`, `interaction.depth`, and `n.trees`. Once the best model is identified, propensity scores are predicted inside `WeightIt`. These can be used inside `WeightIt` to perform IPW or extracted for other implementations. `WeightIt` also supports an offset meaning that logistic regression predictions are supplied to the `GBM` package.

Multiple sources, including package documentation and other research, suggest values for hyperparameters (see McCaffrey, Ridgeway, and Morral 2004; Ridgeway et al. 2024). A very low learning rate, such as 0.01 or 0.0005, allows a smooth descent of the loss function. The model should include a high number of trees, with $10,000$ or $20,000$ being a typical default value. While this may seem excessive, it is required when a low learning rate is used. A grid search process should consider many options including a very high number of trees and even though the optimal model may contain fewer trees. While GBMs often use shallow trees like stumps, allowing a few splits per tree can better model non-linearity and additivity. The package default allows for 3 splits. Based on anecdotal experience, 1 to 5 splits per tree is optimal, consistent with recommendations by McCaffrey, Ridgeway, and Morral (2004).

Another package, `twang`, proves functionality to tune the number of trees, but there are no inbuilt options for tuning of other hyperparameters and so accessory packages such as `caret` must be used. Although `twang` has other useful functionalities which users may wish to implement.

3.1 Hyperparameter Tuning and Workflow

The `WeigthtIt` package seems to have the best options for hyperparameter tuning and integration with a package for assessing balance called `cobalt`. The best information for this package can be found on this website or accessed with `vignette("WeightIt")` inside R after installation using `install.packages("WeightIt")`.

A workflow for hyperparameter tuning in `WeightIt` may be completed as follows:

1. Specify the `criterion` option, which specifies the measure of the *best model*. The available options are the options that the `cobalt` can compute. A simple option to choose may be the average standardised mean difference (SMD) across all covariates called `sdm.mean` or the smallest maximum SDM across covariates called `sdm.max`.
2. Set the number of trees high. The package default is `n.trees = 10000` for binary treatments, but this may be too small depending on the learning rate. Typically, it is best to increase the number of trees to allow slow learners to reach their minimum criterion. There is no modelling downside to a larger number of trees other than computation time as the model will predict propensity scores with a smaller `n.tree` if optimal.
3. Specify the grid search for the depth of the tree called `interaction.depth` and the learning rate called `shrinkage`. These values can be specified using `c()` such as `shrinkage = c(0.0005, 0.001, 0.05, 0.1, 0.2, 0.3)` or as integers such as `interaction.depth = 1:5`. These particular values are heuristically selected *suggestions* of good starting values. Additionally, an offset can be considered by performing a grid search across `offset = c(TRUE,FALSE)`.
4. The model is fit and a grid search is performed. The tune grid and balance statistics can be retrieved with `my_weightit_object$info$best.tune`.
5. The best model should be inspected and to determine if the initial grid is appropriate. If the selection of the best model is at the boundary of a grid search, then a new grid should be created and step 3 and 4 are repeated. For example, if the initial fit is completed with `interaction.depth = 1:5` and the best fit is 5, then a new search can consider `interaction.depth = 3:7` so that the local area around 5 can be searched.
6. Experiment with `bag.fraction`, which means each tree will consider a drawn proportion of observations equal to `bag.fraction`. Iteratively changing `bag.fraction` and assessing balance at each value should be practical. Consider 0.5, 0.67, and 1.
7. Assess balance of covariates and model fit. Covariate balance can be assessed with a balance table or visualisation of the variables using `love.plot()` such as Figure 4.1.
8. The tuning process is stated and reported. Balance tables are presented and discussed. Comparison to other methods of estimation if relevant.
9. Estimation and reporting of treatment effect.

3.2 Example: NSW Jobs Dataset Using R

For demonstration, propensity scores are estimated following the workflow discussed in Section 3.1 to estimate inverse propensity weights (IPW). The NSW jobs dataset arises from a randomised setting as described in Section 6.1.1. Randomisation should eliminate structural differences between groups, but Rosenbaum and Rubin (1983) notes that randomisation only addresses structural balance and does not account for chance imbalance. To address this, propensity scores can mitigate any remaining chance imbalance, providing a more accurate estimate of the treatment effect. This example will include the fitting process of a GBM using `WeightIt` and a logistic regression model using `glm()`. Additionally, balance statistics will be computed leading to a robust estimate of the treatment effect. All code to replicate this process and results is provided.

> **Note 8: Inverse Probability of Treatment Weighting**
>
> Inverse probability of treatment weighting or inverse propensity weighting (IPW) adjusts for confounding in observational data by weighting individuals based on the inverse of their probability of receiving the treatment they actually got. This method creates a *pseudo-population* where treatment assignment is independent of observed covariates, similar to a randomized controlled trial. In this re-weighted population, the treatment and control groups should be have covariate balance, allowing for unbiased estimation of treatment effects. Essentially, IPW simulates random treatment assignment by rebalancing the sample, thereby eliminating confounding and enabling more accurate causal inferences.

3.2.1 Step 1-6: Model Fitting and Tuning

The `glm()` function will fit a conventional propensity score model with logistic regression in R. Logistic regression is performed by specifying the family to be the `binomial()`. Recall the `nsw_formula` is specified in Section 2.2.2.

```
nsw_logit_pmodel <- glm(nsw_formula, data = nsw_data,
                        family=binomial())                    ①

nsw_logit_pscores <- nsw_logit_pmodel$fitted.values           ②
```

① Fits a logistic regression model using the `glm()` function specified to be a logistic model with `family=binomial()` using the previously created `nsw_formula`.
② Extracts the fitted values (propensity scores) from the model.

Using the propensity score column of `nsw_data`, the `WeightIt` package will perform IPW and assign a weight to each observation such that the pseudo-population should exhibit covariate balance. The model object will be called `nsw_logit_weight`.

```
library(WeightIt)
nsw_logit_weight <- weightit(nsw_formula, data = nsw_data,                    ①
                             ps = nsw_logit_pscores,                          ②
                             estimand = "ATE")                                ③
```

① Specifies the formula and data.
② Provides `weightit()` with the propensity scores from the logistic regression function. Note that in practice this can be completed within the `weightit()` function with `method = "glm"`. The separate estimation of the propensity scores is for illustrative purposes.
③ Specifies the estimand as the average treatment effect or ATE. For the purposes of demonstration, this is an arbitrary choice.

A GBM model for propensity scores can be specified using `method = "gbm"` inside the `weightit()` function. To ensure consistent results, running `set.seed(88)` will ensure each tree uses the same `seed` if `bag.fraction` less than 1. The model is fit using the heuristically suggested starting values. Note that this model may take approximately 30 second to fit as a grid search procedure is computationally intensive. Additionally, the best tuning specification is printed to assess if the initial tuning grid is appropriate.

```
set.seed(88)
nsw_boosted_weight <- weightit(nsw_formula, data = nsw_data,                  ①
                               method = "gbm",                               ②
                               estimand = "ATE",
                               shrinkage = c(0.0005, 0.001, 0.05,
                                             0.1, 0.2, 0.3),                 ③
                               interaction.depth = 1:5,
                               bag.fraction = 1,                             ④
                               offset = c(TRUE, FALSE),
                               criterion = "smd.mean",                       ⑤
                               n.trees = 10000)
print(nsw_boosted_weight$info$best.tune)                                      ⑥
```

① Specifies the formula and data.
② Specifies the propensity score prediction method to be a GBM and the estimand to the ATE.
③ Performs a grid search over these values of the learning rate and depth of tree.
④ Requires the model to use every observation in every tree, meaning the model will not perform stochastic gradient boosting. The function will will fit an offset and level GBM and select the specification with the best balance.

⑤ Defines the optimisation criteria to be the tune with the lowest average standardised mean difference (SMD). Additionally, the number of trees will be 10000 which is the package default.
⑥ Prints the tune details of the model with the best covariate balance.

```
  shrinkage interaction.depth distribution use.offset best.smd.mean best.tree
6       0.3                 1    bernoulli      FALSE    0.02253485      2392
```

The best balance across all tuning combinations yields an average SMD of 0.023 showing strong balance compared to the 0.1 threshold. Note averages can conceal extremes and a low average SMD does not mean all variables are balanced. A full balance table is presented in Section 3.2.2 accompanying a discussion of balance.

The best machine has a learning rate of 0.3 and contains 2392 decision stumps (trees with a depth of 1). The learning rate is on the boundary of the initial tuning grid showing that the tuning grid should be re-specified to include values near to 0.3. A reduction in the depth of tree and number of trees will reduce computation time.

The new tune grid will consider `shrinkage = c(0.25, 0.3, 0.35, 0.4, 0.45, 0.5)` as this allows the GBM to consider values between 0.2 and 0.3 and above 0.3 which were missing in the previous grid.

```
   shrinkage interaction.depth distribution use.offset best.smd.mean best.tree
11      0.45                 2    bernoulli      FALSE    0.01965492        95
```

Comparing the two iterations, there is a reduction from 0.022 to 0.02. The optimal tuning values are towards the centre of the tuning grid, implying that an adequate search of the local area has been completed. The best machine has a learning rate of 0.45, a tree depth of 2, and 95 trees. The learning rate is higher than expected, but this also explains why fewer trees are optimal.

Plotting the relationship between the number of trees and the average SMD is informative for the behaviour of the machine. Additionally, Figure 3.1 shows the optimal number of trees is highly variable. If the learning rate is set to `shrinkage = 0.05`, then the best balance is not achieved until near to $20,000$ trees.

For the optimal machine fit, finding that balance worsens as the number of trees increases is just as informative as knowing the correct number of trees. Provided sufficient computational performance, a wide grid search is beneficial in the long run to ensure that each model specification reaches the best balance possible.

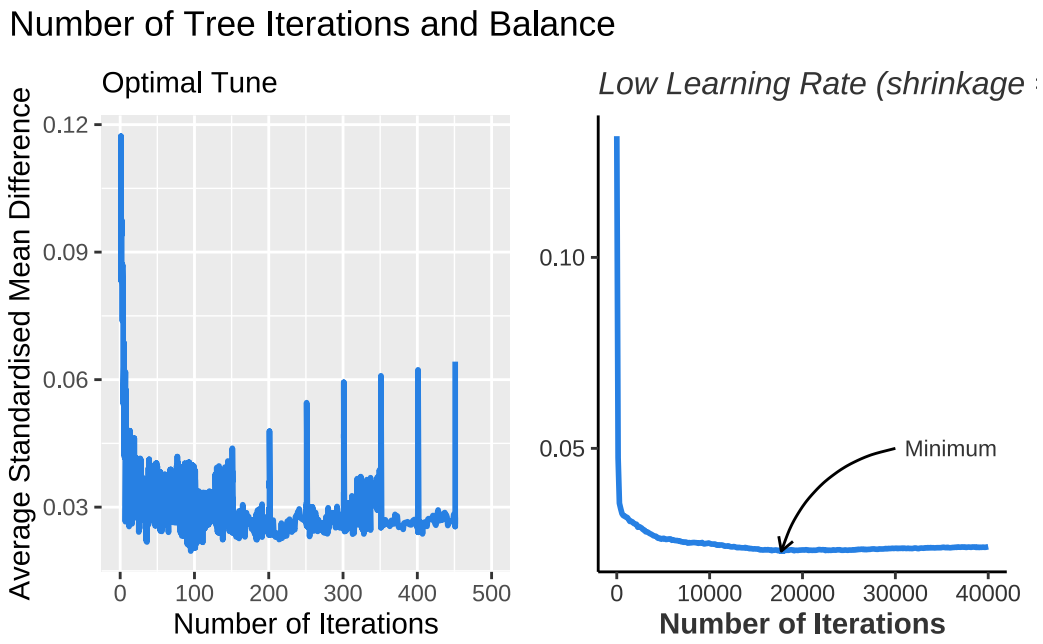


Figure 3.1: Relatoinship between standardised mean difference, number of interations, and learning rate in a GBM model. Please note the difference in horozontal scale between the two learning rates. The model is fit using `weightit` from the `WeightIt` package.

3.2.2 Step 7 and 8: Assessing Balance

> ⚠ The Importance of Discussing Balance
>
> Assessing balance is crucial because it ensures that the treated and control groups are comparable on observed covariates. This comparability is essential for reducing confounding and making valid causal inferences. Without proper balance, differences in outcomes between the groups could be due to pre-existing differences rather than the treatment itself. Balance assessment helps to verify that the propensity score model has effectively adjusted for covariates, creating a pseudo-randomized scenario. This step is vital for the reliability and validity of the study's conclusions. King and Nielsen (2019) notes that many papers that implement propensity score methods do not assess or report a balance in their studies, which can undermine the credibility of the research process and make it hard for readers to understand why results are robust.
> A good resource of information for assessing balance is documentation from the `cobalt` package, which can be viewed by running `vignette("cobalt", package = "cobalt")` in R.

`cobalt` is a powerful package to create tables and visualisations of to assess balance. The package also provides very good integration with other related packages such as `WeightIt` for IPW and `MatchIt` for propensity score matching. Balance tables are created using `bal.tab()`.

```
library(cobalt)                                                          ①
nsw_logit_btab <- bal.tab(nsw_logit_weight,                              ②
                          data = nsw_data,
                          stats = c("mean.diffs","variance.ratios"),     ③
                          binary = "std", continuous = "std",
                          thresholds = c(mean.diffs = 0.1))              ④

nsw_logit_btab <- nsw_logit_btab$Balance[-1,-c(2,3)]                     ⑤
```

① Loads the `cobalt` package. This assumes the package is already installed with `install.packages("cobalt")`
② Uses the `bal.tab()` fucntion to create balance statistics for the previously created `nsw_logit_weight` model.
③ Specifies the calculation of standardised mean differences and variance ratios for each covariate. The mean differences will be standardised for binary and continuous variables.
④ Sets a threshold of balance to be 0.1 to determine if a covariate is balanced.
⑤ Extracts the balance table of the `nsw_logit_btab` object and removes excessive columns. This is only completed for ease of visualisation and is not typically required.

Additionally, `bal.tab()` will create balance tables for the GBM method's IPWs and the raw data. For presentation, `dplyr` combines each of the individual balance tables for presentation

using `kable` and `kableExtra`.

Table 3.1 shows that both logistic regression and the GBM have reduced imbalance. The raw data exhibits imbalance across age, years of education, if someone is gispanic, and if someone has a bachelors degree. Imbalanced datasets leads to biased treatment effect estimation so the estimate of the treatment effect in the raw data may be biased. In this example, logistic regression appears to achieve the best covariate balance although GBM achieves slightly better variance ratios.

3.2.3 Step 9: Results

Finally, the treatment effect can be estimated using `lm_weightit()` from the `WeightIt` package and `avg_comparisons()` from the `marginaleffects` package. `lm_weightit()` fits a linear model with a covariance matrix that accounts for the estimation of weights using IPW. Additionally, `avg_comparisons()` computes the contrast between the treatment and control group to obtain an estimate of the treatment effect.

These steps perform G-computation, meaning that potential outcomes are estimated under treatment and control for each observation (Naimi, Cole, and Kennedy 2017). The contrast of the mean of each of the two potential outcomes is the estimate of the treatment effect. Note that the outcome variable is `re78` which is real income in 1978 meaning that the income is adjusted for inflation. Previously, the treatment indicator was the outcome variable because the propensity scores are a prediction of the treatment indicator.

```
nsw_boosted_lm <- lm_weightit(re78 ~ treat * (age + educ + re75 + black +   ①
                              hisp + degree + marr), data = nsw_data,
                              weights = nsw_boosted_weight$weights)         ②

library(marginaleffects)                                                    ③
nsw_boosted_result <- avg_comparisons(nsw_boosted_lm, variables = "treat")
```

① Uses `lm_weightit()` to compute pseudo-outcomes. The formula here specifies an interaction between the treatment and all other variables. Note that `*` indicates multiplication in R.

② Specifies the `weights` from the `nsw_boosted_weight` object created earlier by the `weightit()` function. Intuitively, this is performing linear regression using the pseudo-population, where the pseudo-population is created weighting the data by `nsw_boosted_weight$weights`.

③ Computes a comparison between the potential outcomes as well as standard errors for inference.

Additionally, this process is followed for the logistic regression propensity scores and the results are combined in to a table for comparison.

Table 3.1: Standardised mean difference (a measure of balance) across different covariates in the National Supported Word dataset. The values are categorised for different propenensity score methods allowing a comparison. Balance tables are computed using `bal.tab()` from the `cobalt` package.

Variable	Type	SMD	Balanced	Variance Ratio
Raw Data				
Age	Contin.	0.1066	No	1.0278
Education	Contin.	0.1281	No	1.5513
Income 1975	Contin.	0.0824	Yes	1.0763
Black	Binary	0.0449	Yes	NA
Hispanic	Binary	-0.2040	No	NA
Degree	Binary	0.2783	No	NA
Married	Binary	0.0902	Yes	NA
IPTW: Logistic Regression				
Age	Contin.	-0.0001	Yes	0.9809
Education	Contin.	0.0012	Yes	1.2725
Income 1975	Contin.	0.0081	Yes	0.7971
Black	Binary	0.0006	Yes	NA
Hispanic	Binary	-0.0031	Yes	NA
Degree	Binary	0.0009	Yes	NA
Married	Binary	0.0045	Yes	NA
IPTW: Boosting				
Age	Contin.	-0.0065	Yes	0.9086
Education	Contin.	0.0220	Yes	1.1391
Income 1975	Contin.	-0.0152	Yes	1.0134
Black	Binary	0.0028	Yes	NA
Hispanic	Binary	-0.0547	Yes	NA
Degree	Binary	0.0481	Yes	NA
Married	Binary	0.0085	Yes	NA

Table 3.2: Comparison of estimates of the average treatment for the National Supported Work data.

	Estimate	SE	P.Value	Lower.CI	Upper.CI
Logistic Regression	1610.786	668.4870	0.0160	300.5756	2920.997
GBM	1609.947	669.4201	0.0162	297.9081	2921.987

Table 3.2 shows that both estimates of the treatment effect are nearly identical at $1610 with logistic regression inferring a $0.86 larger treatment effect. Additionally, these results are statistically significant at the 5% level with nearly identical standard errors.

4 Replication Case Study

The replication study focuses on a paper titled *The Impact of Coffee Certification on Small-Scale Producers' Livelihoods: A Case Study from the Jimma Zone, Ethiopia* published in Agricultural Economics (2012) by Pradyot Ranjan Jena, Bezawit Beyene Chichaibelu, Till Stellmacher, and Ulrike Grote. This paper explores the effects of coffee certification schemes on the economic wellbeing of small-scale coffee farmers in Ethiopia, particularly examining whether these schemes contribute to poverty reduction and improved livelihoods among smallholders.

The central theme of the paper is the evaluation of certification schemes, such as Fairtrade and organic certification, as tools for enhancing the income stability and economic resilience of small-scale coffee producers. Certification is seen as a potential tool for economic growth and and environmental sustainability and so it is important to understand the impact on small-scale farmers. Table 4.1 summarises the variables used in the propensity score model.

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union


Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

Table 4.1: List of covaraites in Jena et al. (2012). Per capita income is the outcome and certification is the treatment. Other covariates are used in the propensity score model.

Name	Codename	Description
Per Capita Income	`percapitaincome_daygrosseq`	Average income earned per person within a farming household
Certification (Treatment/ Control)	`certified`	If the farming household is certified (=1) or otherwise (=0)
Household Age	`age_hh`	Age of the head of the household in years
Squared Household Age	`agesq`	Age of the head of the household squared
Gender	`gender`	Gender of the head of household (male = 1 and female = 0)
Dependency Ratio	`depratio`	Household members below 14 and above 65 years divided by rest of the household member
Education Level	`edu`	Education of the head of household in years
Years of Coffee Production	`years_cofeproduction`	Years of experience in coffee farming
Log Total Land	`logtotal_land`	Logarithm of total land size in hectares
Access to Credit	`access_credit`	Household has access to credit (yes = 1, otherwise = 0)
Bad Weather last year (2008–2009)	`badweat`	If the household was affected by floods/droughts during the
Non-farm Income Access	`nonfarmincome_access`	If the household has access to nonfarm income

Jena et al. (2012) define livelihood as a combination of per capita income, total income, per capita consumption, and yield per hectare. For simplicity, this replication will only use per capita income as a dependent variable. This measure is selected as per capita income is a direct measurement of the income of those potentially impacted by certification. Additionally, the replication uses the same variables as the original paper so any difference in estimates or covariate balance can be attributed to the propensity score model.

Randomisation into certified and uncertified is not possible and it is likely that farmers who seek certification are different than farmers who don't. Thus, there is selection bias leading to structural differences between groups so a contrast in means between the certified (treated) and

uncertified (control) farmers would be biased. Propensity scores are used to create covariate balance and reduce bias of the estimated treatment effect. The paper did not assess the balance of covariates. However, this provides a good opportunity to assess covariate balance in the initial paper and the repeat the analysis using a machine learning propensity model.

4.1 Replication of Original Results

Jena et al. (2012) provides a replication package including Stata code that uses Stata's psmatch2 package to perform nearest neighbour matching with replacement and common support trimming. Common support trimming means that any observations outside the commonly overlapping are are discarded. The results of the paper are be fully replicated using the `MatchIt` package inside R.

Table 4.2: Replication of results in Jena et al. (2012). Note a slight difference in standard error which should be 1.1 as the `MatchItSE` package uses a trivially different method than psmatch2 in Stata (see Abadie and Imbens 2006). Matching is performed by `matchit()` from the `MatchIt` package.

	Estimate	SE	P.Value	Lower.CI	Upper.CI
Replicated Result	-0.1538	0.9898	0.835	-1.6009	1.2934

Table 4.2 shows the replicated result obtained by Jena et al. (2012). The intriguing finding of the paper is that the average treatment effect on the treated (ATT) is negative. That is, of the farmers that become certified, their per capita income is expected to decrease by $0.15 per day. Intuition and proponents of certification schemes suggest that certification leads to an increase of income. If certification negatively impacts income, it would call into question a significant effort to engage in certification and fair trade practices.

Jena et al. (2012) does not perform any discussion or consideration of balance in their paper and so it is unclear if propensity score matching results in covariate balance. The `cobalt` package creates balance tables using `bal.tab()` and a visualisation using `love.plot()`.

Table 4.3: The standardised mean difference (SMD) for each covariate in Jena et al. (2012) using a logistic regression propensity model and propensity score matching. Across each of the covariates, a balance threshold is set at 0.1 to indicate if a covariate is balanced. Binary and continuous variables are both standardised over the treatment group. SMDs are computed using `bal.tab()` from the `cobalt` package.

	Type	Diff.Adj	M.Threshold	V.Ratio.Adj
Household Age	Contin.	-0.2723	No	1.0728

Table 4.3: The standardised mean difference (SMD) for each covariate in Jena et al. (2012) using a logistic regression propensity model and propensity score matching. Across each of the covariates, a balance threshold is set at 0.1 to indicate if a covariate is balanced. Binary and continuous variables are both standardised over the treatment group. SMDs are computed using `bal.tab()` from the `cobalt` package.

	Type	Diff.Adj	M.Threshold	V.Ratio.Adj
Squared Household Age	Contin.	-0.2552	No	1.1430
Non-farm Income Access	Binary	0.3005	No	NA
Log Total Land	Contin.	0.2597	No	1.2969
Dependency Ratio	Contin.	-0.3996	No	0.9789
Bad Weather	Binary	0.2016	No	NA
Education Level	Contin.	0.2443	No	1.0338
Gender	Binary	-0.1324	No	NA
Years of Coffee Production	Contin.	-0.3400	No	0.9111
Access to Credit	Binary	0.1949	No	NA

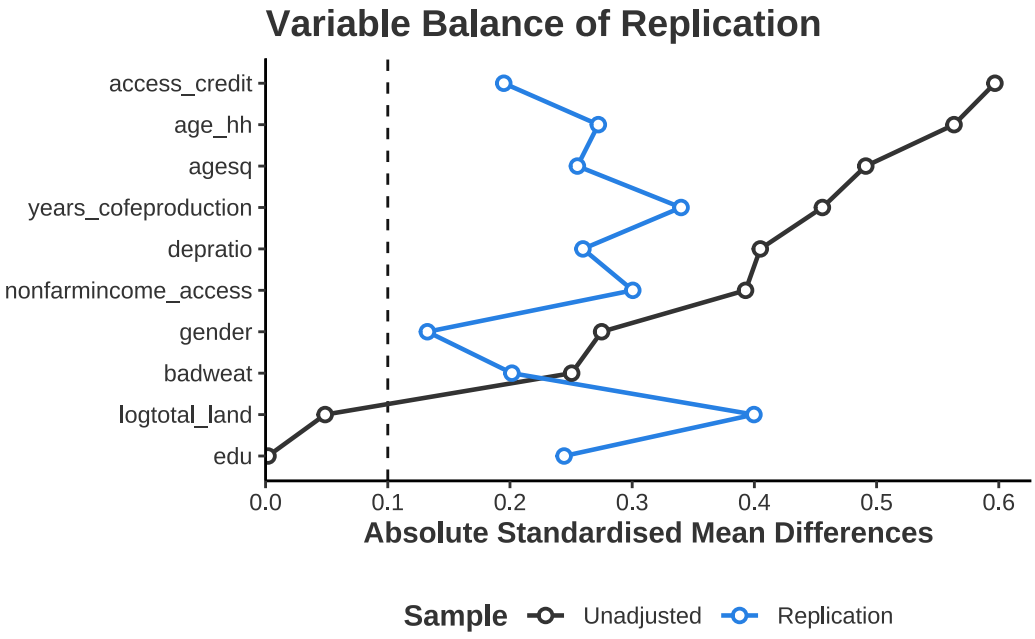


Figure 4.1: A visual representation of Table 4.3 called a love plot. Additionally, the unadjusted (raw data) SMDs are displayed for comparison. Variables are ordered by the SMD in the unadjusted data. Plot is created using `love.plot()` from the `cobalt` package.

Table 4.3 and Figure 4.1 show that propensity score matching has obtained very poor balance. Based on the 10% rule discussed in Section 2.1.1, not a single variable is balanced and so the estimate of the treatment effect is likely to be biased by structural differences between control and certified.

Four key variables: *age*, *gender*, *education*, and *access to credit* all exhibit poor balance. These variables are strong confounders in theory and so emphasising balance in these variables is critical to making a robust causal inference. Perhaps there is gender or age discrimination in the certification process. Perhaps, those with lesser education may struggle to obtain certification. Perhaps those who have less access to credit are unable to afford to become certified. Moving forward, these variables must exhibit better covariate balance to make a robust conclusion.

Figure 4.2 shows the effect of common support trimming. Table 4.4 shows 34 total observations are dropped of which 33 are treated and 1 are control. By dropping these observations, PSM avoids making poor matching which should lead to better covariate balance. When observations are discarded, the estimand is no longer the ATT. Instead, it is refereed to as the average treatment effect on the matched or ATM. There is a significant reduction in the effective sample size in the control group from 82 to 21 individuals.

Table 4.4: The effective sample size resulting from the use of propensity score matching in Jena et al. (2012). The effective sample size (ESS) is displayed in the unweighted (raw) and matched data as well as the number of disarded observations. Computed using `bal.tab()` from the `cobalt` package.

	Control	Certified
All (ESS)	82	164
All (Unweighted)	82	164
Matched (ESS)	21	131
Matched (Unweighted)	42	131
Unmatched	39	0
Discarded	1	33

Overall, the propensity score matching in Jena et al. (2012) is poor and results in unbalanced covariates and a loss of estimand.

4.2 Further Modelling

To improve the poor balance achieved by the Jena et al. (2012), there are two strategies to obtain better balance. First, the propensity scores can be re-estimated using machine learning to obtain better calibrated propensity scores. Second, inverse propensity weighting (IPW) can

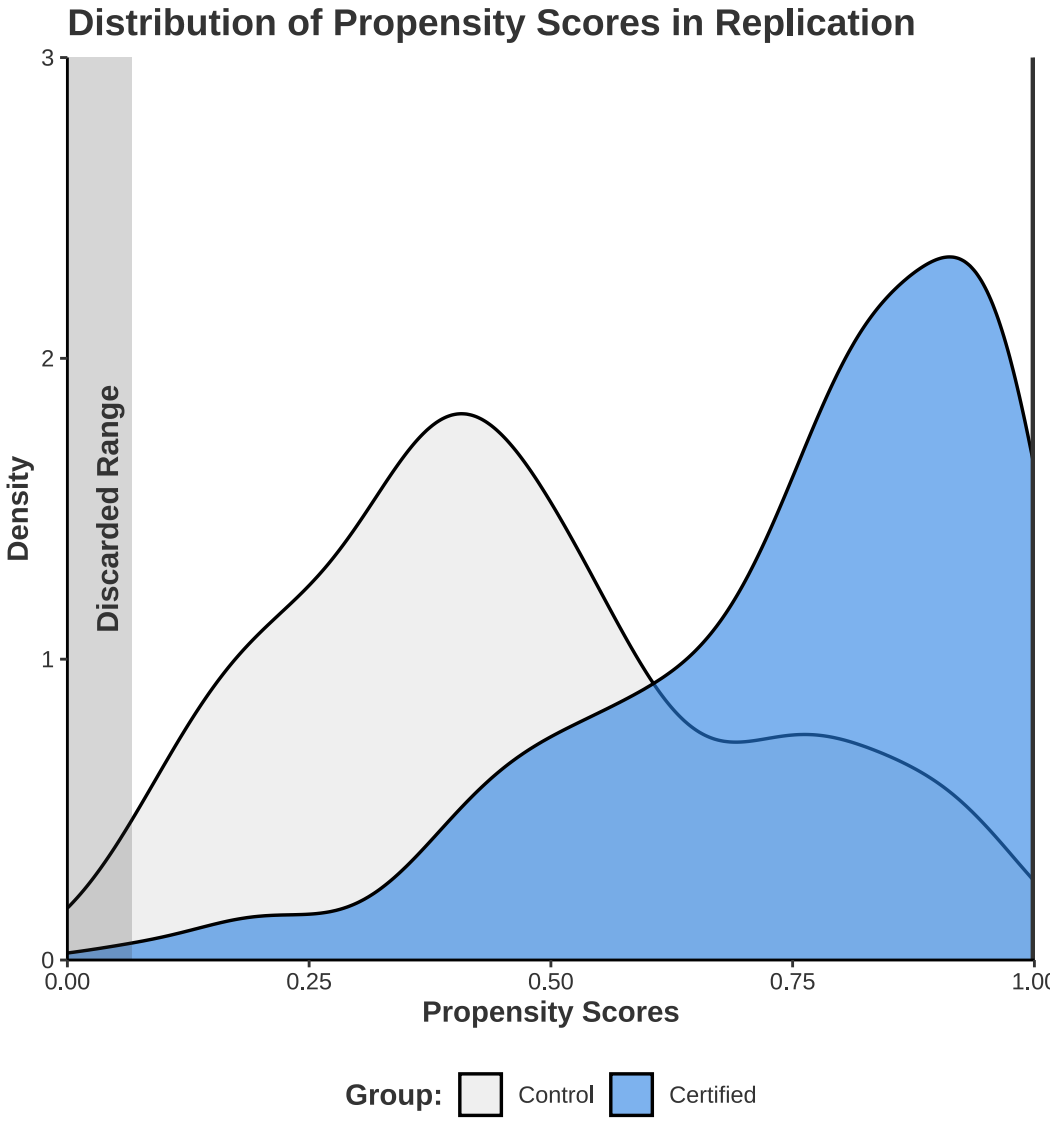


Figure 4.2: Density estimates of the propensity scores from Jena et al. (2012) using logistic regression. Propensity score matching discards some observations as displayed by the discarded range on the left. A single observation is discarded on the right.

be used instead of propensity score matching (PSM). IPW should ensure that the sample size remains the same as no observations are lost through a matching process. IPW should retain all observations and preserve the estimand as the ATT. Additionally, IPW is generally more efficient as a pseudo-population is based on precise weights compared to matched observations that are based on approximate similarity.

The machine learning propensity scores will be estimated using the `WeightIt` package in the same process as Chapter 3. To select the tuning criteria, consider that Figure 4.1 that shows a significant range of balance levels in the raw data. Knowing this, the model is tuned using `criterion = "smd.max"` as reducing the priority is to reduce the extremely unbalanced covariates even if this leads to a higher average SMD.

> **Discussion of Tuning**
>
> Initially, a tuning grid considering shrinkage values of $0.001, 0.005, .01, 0.05, 0.1,$ and 0.2 were considered using 10000 trees with a depth between 1 and 5. The best tuning performance was found with shrinkage of 0.2 and 9 trees which were three splits 3 deep. As such, the tuning grid was redefined in a second iteration to use $0.1, 0.15, 0.2, 0.25, 0.3, 0.35,$ and 0.4 with only 1000 trees with between 2 and 5 depth. The second fit, suggested a learning rate of 0.35 so the local area of $0.3, 0.325, 0.350, 0.375,$ and 0.4 is searched in the final fit.

Of course there is no guarantee that the GBM model will perform the best and so a logistic model is also fitted. An interesting comparison is between the SMDs in the matched data and in the weighted sample. Any differences between the two samples relates to the difference between PSM and IPW as the propensity scores are identical.

4.3 Comparison of Methods

In some of the earlier code chunks, the `cobalt` package's `bal.tab()` computed balance tables which are combined together in Table 4.5 to provide a comparison between methods. For a visual interpretation, `love.plot()` creates Figure 4.3.

```
Attaching package: 'data.table'

The following objects are masked from 'package:dplyr':

    between, first, last
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
v readr     2.1.5
-- Conflicts ---------------------------------------- tidyverse_conflicts() --
x data.table::between()    masks dplyr::between()
x dplyr::filter()          masks stats::filter()
x data.table::first()      masks dplyr::first()
x kableExtra::group_rows() masks dplyr::group_rows()
x lubridate::hour()        masks data.table::hour()
x lubridate::isoweek()     masks data.table::isoweek()
x dplyr::lag()             masks stats::lag()
x data.table::last()       masks dplyr::last()
x lubridate::mday()        masks data.table::mday()
x lubridate::minute()      masks data.table::minute()
x lubridate::month()       masks data.table::month()
x lubridate::quarter()     masks data.table::quarter()
x lubridate::second()      masks data.table::second()
x purrr::transpose()       masks data.table::transpose()
x lubridate::wday()        masks data.table::wday()
x lubridate::week()        masks data.table::week()
x lubridate::yday()        masks data.table::yday()
x lubridate::year()        masks data.table::year()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

There are three notable findings:

1. PSM has performed very poorly relative to IPW even when matching dropps a significant number of observations.
2. A GBM model has resulted in better covariate balance than logistic regression for most covariates. Using a 10% threshold for determining balance, logistic regression leaves 5 variables unbalanced and the GBM leaves 3 variables unbalanced. Additionally, the degree of unbalance is larger for logistic regression.
3. Logistic regress has a satisfactory average SMD of 0.0768603. Boosting has an average SMD of 0.0498114 which is excellent and meets a rigorous threshold of 5%.
4. The covariate with the highest SMD is *household age* (0.245%) in logistic regression and *bad weather* (0.191) in the GBM.

Table 4.5: Comparison of standardised mean difference (SMD) using different propensity score models. Across each of the covariates, a balance threshold is set at 0.1 to indicate if a covariate is balanced. Binary and continuous variables are both standardised over the treatment group. SMDs are computed using `bal.tab()` from the `cobalt` package.

Variable	Type	SMD	Balance Threshold	Variance Ratio
Raw Data				
Household Age	Contin.	0.5634	No	0.8650
Squared Household Age	Contin.	0.4912	No	1.0070
Non-farm Income Access	Binary	0.3928	No	NA
Log Total Land	Contin.	-0.4048	No	0.5507
Dependency Ratio	Contin.	0.0487	Yes	1.2371
Bad Weather	Binary	0.2505	No	NA
Education Level	Contin.	-0.0020	Yes	0.7272
Gender	Binary	0.2750	No	NA
Years of Coffee Production	Contin.	0.4557	No	1.3621
Access to Credit	Binary	0.5968	No	NA
Logistic Regression and IPTW				
Household Age	Contin.	0.2449	No	0.9275
Squared Household Age	Contin.	0.2277	No	1.0724
Non-farm Income Access	Binary	0.1701	No	NA
Log Total Land	Contin.	-0.0923	Yes	0.8564
Dependency Ratio	Contin.	0.1138	No	1.3877
Bad Weather	Binary	0.1941	No	NA
Education Level	Contin.	0.0473	Yes	0.9217
Gender	Binary	-0.0465	Yes	NA
Years of Coffee Production	Contin.	-0.0613	Yes	1.1125
Access to Credit	Binary	-0.0292	Yes	NA
Boosted Machine with IPTW				
Household Age	Contin.	0.0669	Yes	1.2690
Squared Household Age	Contin.	0.0989	Yes	1.4911
Non-farm Income Access	Binary	0.0579	Yes	NA
Log Total Land	Contin.	-0.0284	Yes	0.8760
Dependency Ratio	Contin.	-0.0623	Yes	0.7660
Bad Weather	Binary	0.1915	No	NA
Education Level	Contin.	0.1377	No	1.0735
Gender	Binary	-	Yes	NA

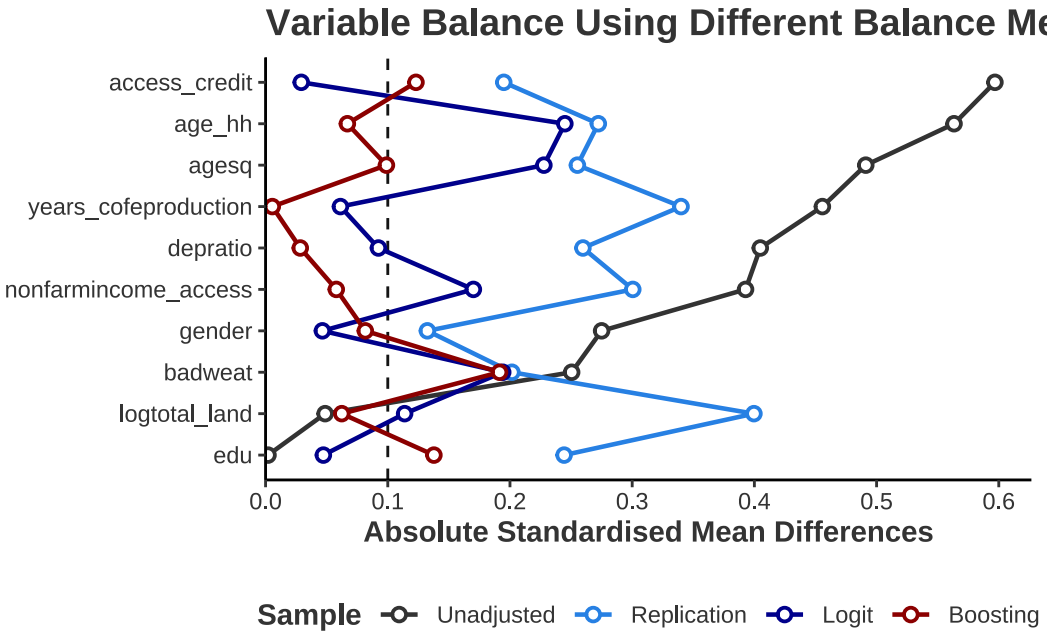


Figure 4.3: Visual representation of Table 4.5 called a love plot and displays the standardised mean difference (SMD) of covariates in Jena et al. (2012). Additionally, the unadjusted SMDs are displayed for comparison. Variables are ordered by the SMD in the unadjusted data.

4.4 Results

Now that satisfactory covariate balance is achieved, the treatment effect can be estimated under logistic regression, the GBM, and then compared to the result in the paper. Note that the estimand in the paper is intended to be the average treatment effect (ATT) but dropped observations mean the actual treatment effect is the average treatment effect on matched (ATM) individuals. In theory, better covariate balance should lead to a better estimate of the ATT so a comparison of the estimates is interesting. As in Section 3.2.3, the results will be completed using G-computation with the `lm_weightit()` and `avg_comparisons()` functions.

Table 4.6: Estimates of the average treatment effect on the treated of certification on per capita income across different propensity score models and methods. Created using `WeightIt`, `MatchIt`, and `Cobalt` packages.

	Estimate	SE	P.Value	Lower.CI	Upper.CI
Rep. Result (Logistic with PSM)	-0.1538	0.9898	0.8350	-1.6009	1.2934
Logistic Regression and IPW	-1.5824	0.6072	0.0092	-2.7724	-0.3924
Generalized Boosting Machine and IPW	-1.0187	0.5196	0.0499	-2.0372	-0.0003

Table 4.6 shows the estimates of the treatment effect across different methods. Recall that Jena et al. (2012) estimate a an effect of -0.15 implying that daily income reduces by 0.15 if a farmer becomes certified. This result is not statistically significant.

The IPW estimate is -1.58 implying that certification leads to a \$1.58 decrease in daily income. This coefficient is much larger than than the original paper by a magnitude of 10. Additionally, this estimate is statistically significant at the 1% level. The GBM estimate is -1.02 which predicts a decrease in daily income by \$1.02 when a farmer becomes certified. This finding is statistically significant at the 5% level.

The reason for a large difference is threefold. First, better covariate balance by using a GBM and IPW and should result in a more robust estimate. Of course better covariate balance alone does not guarantee robust results but it is a step in the right direction. Theoretically, weighting on the inverse of the propensity scores from a GBM results in the best estimate so the paper may significantly underestimate the coefficient. Second, a different estimand will often lead to a different estimate of the treatment effect. It is not clear or directly estimable how much of the difference in estimate results from switching from the ATM to the ATT. Related to this, the data used to estimate the treatment effect is different as there are no dropped observations either of the IPW estimates. Third…

The most interesting result is that the estimates become even more negative. One may expect that the result from a better balanced sample would become positive to align with theoret-

ical motivations for cerification policies. Jena et al. (2012) presented two explanations for why certification shows no positive impact. Firstly, the authors note that the prices offered by certified cooperatives are not significantly different from those provided by non-certified cooperatives. Secondly, a substantial portion of coffee—about 75%—is sold to private traders, who often pay higher prices to non-certified farmers. Additionally, from qualitative interviews with farmers, the authors note that policies and arrangements within different cooperatives exhibit heterogeneity so the impact of certification may relate more to the structure of the cooperatives.

An additional answer is the impact of reverse causality. A general problem is causal inference is that the direction of causality is not always known. While it is most intuitive that coffee certification would impact income, it is also possible that a farmers daily income might determine their certification. Suppose that proponents of fairtrade and certification are correct that it will increase income and benefit livelihood. If farmers are are of this, then perhaps the lowest income farmers are most likely to attempt to become certified to increase their income. Additionally, income likely has a reverse causal relationship with many of the explanatory variables. For example, a higher income may lead to better access to credit and the accumulation of land.

In summary, the analysis demonstrates that using more advanced methods like GBM and IPW not only improves covariate balance but also leads to significantly larger and more negative estimates of the treatment effect compared to the original study. This suggests that previous estimates may have underestimated the negative impact of certification on daily income. The findings highlight the importance of methodological rigor in estimating causal effects and raise critical questions about the broader implications of certification policies, particularly when considering potential reverse causality and the varying structures of cooperatives. This analysis underscores the need for careful interpretation of treatment effects, especially in policy-relevant research.

5 Conclusion and Summary

References

Abadie, Alberto, and Guido W. Imbens. 2006. "Large sample properties of matching estimators for average treatment effects." *Econometrica* 74 (1): 235–67. https://doi.org/10.1111/j.1468-0262.2006.00655.x.

Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2024. *rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown.

Arel-Bundock, Vincent, Noah Greifer, and Andrew Heiss. Forthcoming. "How to Interpret Statistical Models Using marginaleffects in R and Python." *Journal of Statistical Software*, Forthcoming.

Austin, Peter. 2011. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." *Multivariate Behavioral Research* 46 (3): 399–424. https://doi.org/10.1080/00273171.2011.568786.

Austin, Peter C. 2008. "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003." *Statistics in Medicine* 27 (April): 2037–49. https://doi.org/10.1002/sim.3150.

Bader-El-Den, Mohammed, Eleman Teitei, and Todd Perry. 2019. "Biased Random Forest for Dealing with the Class Imbalance Problem." *IEEE Transactions on Neural Networks and Learning Systems* 30 (7): 2163–72. https://doi.org/10.1109/TNNLS.2018.2878400.

Barrett, Tyson, Matt Dowle, Arun Srinivasan, Jan Gorecki, Michael Chirico, and Toby Hocking. 2024. *data.table: Extension of "data.frame"*. https://CRAN.R-project.org/package=data.table.

Breiman, Leo. 1996. "Bagging predictors." *Machine Learning* 24: 123–40. https://doi.org/10.3390/risks8030083.

Breiman, L, Jerome H Friedman, Richard A Olshen, and C J Stone. 1984. "Classification and Regression Trees." *Biometrics* 40: 874. https://api.semanticscholar.org/CorpusID:29458883.

Brieman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324.

Brookhart, M. Alan, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Stürmer. 2006. "Variable selection for propensity score models." *American Journal of Epidemiology* 163 (12): 1149–56. https://doi.org/10.1093/aje/kwj149.

Cannas, Massimo, and Bruno Arpino. 2019. "A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting." *Biometrical Journal* 61 (4): 1049–72. https://doi.org/10.1002/bimj.201800132.

Cunningham, Scott. 2021. "Matching and Subclassification." In *Causal Inference: The Mixtape*, 175–240. Yale University Press. https://doi.org/10.2307/j.ctv1c29t27.8.

Ferri-García, Ramón, and María Del Mar Rueda. 2020. "Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys." *PLoS ONE* 15 (4): 1–19. https://doi.org/10.1371/journal.pone.0231500.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29 (5): 1189–1232. https://www.jstor.org/stable/2699986.

Goller, Daniel, Michael Lechner, Andreas Moczall, and Joachim Wolff. 2020. "Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed." *Labour Economics* 65 (March). https://doi.org/10.1016/j.labeco.2020.101855.

Greifer, Noah. 2024a. *cobalt: Covariate Balance Tables and Plots*. https://CRAN.R-project.org/package=cobalt.

———. 2024b. *WeightIt: Weighting for Covariate Balance in Observational Studies*. https://CRAN.R-project.org/package=WeightIt.

Heinrich, Carolyn. 2010. "A Primer for Applying Propensity-Score Matching." *Development*, no. August: 59. http://www.iadb.org/document.cfm?id=35320229.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software* 42 (8): 1–28. https://doi.org/10.18637/jss.v042.i08.

Huntington-Klein, Nick, and Malcolm Barrett. 2021. *causaldata: Example Data Sets for Causal Inference Textbooks*. https://CRAN.R-project.org/package=causaldata.

Jena, Pradyot Ranjan, Bezawit Beyene Chichaibelu, Till Stellmacher, and Ulrike Grote. 2012. "The impact of coffee certification on small-scale producers' livelihoods: A case study from the Jimma Zone, Ethiopia." *Agricultural Economics (United Kingdom)* 43 (4): 429–40. https://doi.org/10.1111/j.1574-0862.2012.00594.x.

King, Gary, and Richard Nielsen. 2019. "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27 (4): 435–54. https://doi.org/10.1017/pan.2019.11.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart. 2010. "Improving propensity score weighting using machine learning." *Statistics in Medicine* 29: 337–46. https://doi.org/10.1002/sim.3782.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. 2004. "Propensity score estimation with boosted regression for evaluating causal effects in observational studies." *Psychological Methods* 9 (4): 403–25. https://doi.org/10.1037/1082-989X.9.4.403.

Naimi, Ashley I., Stephen R. Cole, and Edward H. Kennedy. 2017. "An introduction to g methods." *International Journal of Epidemiology* 46 (2): 756–62. https://doi.org/10.1093/ije/dyw323.

Olson, Matthew A., and Abraham J. Wyner. 2018. "Making Sense of Random Forest Probabilities: a Kernel Perspective," 1–35. http://arxiv.org/abs/1812.05792.

Pedersen, Thomas Lin. 2024. *patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

Qiu, Yixuan, and authors/contributors of the included software. See file AUTHORS for details. 2024. *showtext: Using Fonts More Easily in r Graphs*. https://CRAN.R-project.org/package=showtext.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ridgeway, Greg, and GBM Developers. 2024. *gbm: Generalized Boosted Regression Models*. https://CRAN.R-project.org/package=gbm.

Ridgeway, Greg, Dan Mccaffrey, Andrew Morral, Matthew Cefalu, Lane Burgette, and Beth Ann Griffin. 2024. "Toolkit for Weighting and Analysis of Nonequivalent Groups: A Tutorial for the R TWANG Package." https://doi.org/10.7249/tl136.1.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70 (1): 41–55. https://doi.org/10.1017/CBO9780511810725.016.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Experimental and Observational Studies." *Journal of Educational Psychology* 66 (5): 688–701. https://doi.org/10.1002/j.2333-8504.1972.tb00631.x.

Schuster, Tibor, Wilfrid Kouokam Lowe, and Robert W. Platt. 2016. "Propensity score model overfitting led to inflated variance of estimated odds ratios." *Journal of Clinical Epidemiology* 80: 97–106. https://doi.org/10.1016/j.jclinepi.2016.05.017.

Setoguchi, Soko, Sebastian Schneeweiss, Alan M. Brookhart, Robert J. Glynn, and Francis E. Cook. 2008. "Evaluating uses of data mining techniques in propensity score estimation: a simulation study." *Pharmacoepidemiology and Drug Safety* 17 (March): 546–55. https://doi.org/10.1002/pds.

Smith, Jeffrey A., and Petra E. Todd. 2005. *Does matching overcome LaLonde's critique of nonexperimental estimators?* Vol. 125. 1-2 SPEC. ISS. https://doi.org/10.1016/j.jeconom.2004.04.011.

Splawa-Neyman, Jerzy. 1923. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." PhD thesis. https://doi.org/10.1214/ss/1177012031.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society* 58 (1): 267–88. https://www.jstor.org/stable/pdf/2346178.pdf?refreqid=fastly-default%7B/%%7D3Aff57285d6b8854126d21a135984fc4ca%7B/&%7Dab%7B/_%7Dsegments=%7B/&%7Dorigin=%7B/&%7Dinitiator=%7B/&%7DacceptTC=1.

Tu, Chunhao. 2019. "Comparison of various machine learning algorithms for estimating generalized propensity score." *Journal of Statistical Computation and Simulation* 89 (4): 708–19. https://doi.org/10.1080/00949655.2019.1571059.

Ushey, Kevin, and Hadley Wickham. 2024. *renv: Project Environments*. https://CRAN.R-project.org/package=renv.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Im-*

plementing Reproducible Computational Research, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2024. *knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with "kable" and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

R Version Control

Please note that future updates to these packages may impact replication of results.

Package	Version	Citation
base	4.4.1	R Core Team (2024)
causaldata	0.1.3	Huntington-Klein and Barrett (2021)
cobalt	4.5.5	Greifer (2024a)
data.table	1.15.4	Barrett et al. (2024)
gbm	2.2.2	Ridgeway and Developers (2024)
kableExtra	1.4.0	Zhu (2024)
knitr	1.48.1	Xie (2014); Xie (2015); Xie (2024)
marginaleffects	0.21.0	Arel-Bundock, Greifer, and Heiss (Forthcoming)
MatchIt	4.5.5	Ho et al. (2011)
MatchItSE	1.0	(**MatchItSE?**)
patchwork	1.2.0	Pedersen (2024)
randomForest	4.7.1.1	Liaw and Wiener (2002)
renv	1.0.7	Ushey and Wickham (2024)
rmarkdown	2.27	Xie, Allaire, and Grolemund (2018); Xie, Dervieux, and Riederer (2020); Allaire et al. (2024)
showtext	0.9.7	Qiu and See file AUTHORS for details. (2024)
tidyverse	2.0.0	Wickham et al. (2019)
WeightIt	1.2.0	Greifer (2024b)

6 Appendix

6.1 Datasets

6.1.1 National Supported Work Data

```
library('causaldata')
data("nsw_mixtape", package = "causaldata")
nsw_data <- as.data.frame(nsw_mixtape)

nsw_data$data_id <- seq(1,length(nsw_data$data_id))

nsw_data$degree <- abs(nsw_data$nodegree-1)

nsw_data$nodegree <- NULL
```

6.1.2 Coffee Data from Jena et al. (2012)

```
library(haven)
coffee_data <-  read_dta("C:/Users/mitch/OneDrive - University of Otago/Mitchell Research/Pr

coffee_data <- zap_formats(coffee_data)

coffee_data <- coffee_data[-c(56,84,156 ),]
```

6.2 Functions

```
extract_comparison_results <- function(results) {
  extracted_results <- data.frame(
    Estimate = results$estimate,
```

```
    SE = results$std.error,
    P.Value = results$p.value,
    Lower.CI = results$conf.low,
    Upper.CI = results$conf.high
  )

  return(extracted_results)
}
```

6.3 Theming

```
library(ggplot2)
library(showtext)

font_add_google(name = "Source Sans Pro", family = "Source Sans Pro")

showtext_auto()

custom_ggplot_theme <- theme_classic(base_size = 11,
                                     base_family = "Source Sans Pro") +
    theme(
      text = element_text(color = "#333333"),
      plot.background = element_blank(),
      panel.background = element_blank(),
      axis.text = element_text(color = "#333333"),
      axis.title = element_text(color = "#333333", face = "bold"),
      legend.text = element_text(color = "#333333"),
      legend.title = element_text(color = "#333333", face = "bold",),
      plot.title = element_text(size = 14, face = "bold", color = "#333333"),
      plot.subtitle = element_text(size = 12, face = "italic",
                                   color = "#333333"),
      strip.text = element_text(face = "bold", family = "Source Sans Pro",
                                color = "#333333"), legend.position="bottom")

theme_set(custom_ggplot_theme)
```

6.4 Code Provided for PDF Output

```
load(file = "globals.RData")
library('causaldata')
data("nsw_mixtape", package = "causaldata")
nsw_data <- as.data.frame(nsw_mixtape)

nsw_data$data_id <- seq(1,length(nsw_data$data_id))

nsw_data$degree <- abs(nsw_data$nodegree-1)

nsw_data$nodegree <- NULL
library(haven)
coffee_data <-  read_dta("C:/Users/mitch/OneDrive - University of Otago/Mitchell Research/Pr

coffee_data <- zap_formats(coffee_data)

coffee_data <- coffee_data[-c(56,84,156 ),]
extract_comparison_results <- function(results) {
  extracted_results <- data.frame(
    Estimate = results$estimate,
    SE = results$std.error,
    P.Value = results$p.value,
    Lower.CI = results$conf.low,
    Upper.CI = results$conf.high
  )

  return(extracted_results)
}
library(ggplot2)
library(showtext)

font_add_google(name = "Source Sans Pro", family = "Source Sans Pro")

showtext_auto()

custom_ggplot_theme <- theme_classic(base_size = 11,
                                     base_family = "Source Sans Pro") +
    theme(
      text = element_text(color = "#333333"),
      plot.background = element_blank(),
```

```
      panel.background = element_blank(),
      axis.text = element_text(color = "#333333"),
      axis.title = element_text(color = "#333333", face = "bold"),
      legend.text = element_text(color = "#333333"),
      legend.title = element_text(color = "#333333", face = "bold",),
      plot.title = element_text(size = 14, face = "bold", color = "#333333"),
      plot.subtitle = element_text(size = 12, face = "italic",
                                   color = "#333333"),
      strip.text = element_text(face = "bold", family = "Source Sans Pro",
                                color = "#333333"), legend.position="bottom")

theme_set(custom_ggplot_theme)
save.image(file = "globals.RData")
```