18

On Machine Learning Methods for Propensity Score Matching and Weighting in Educational Data Mining Applications

Juanjuan Fan, Joshua Beemer, Xi Yan, and Richard A. Levine

CONTENTS

18.1 Introduction

A well-designed randomized experiment can yield unbiased treatment effect because the randomized treatment allocation balances the baseline covariates between treated and control subjects. However, a randomized trial is not always feasible due to ethical or practical reasons. The observational study, as an alternative to the randomized experiment, can not only avoid the possible moral hazard (for example, when the treatment of interest is smoking) but also be less expensive. However, the observational study may produce biased estimates of the treatment effect due to treatment of self-selection. Observational studies are at the heart of data analytics for institutional and student success research since students are often allowed to decide for themselves whether or not to take part in educational interventions.

Propensity score (Rosenbaum & Rubin, 1983), defined as the probability of being treated conditional on observed covariates, is a useful tool for deriving unbiased estimates of the treatment effect based on observational study data. Subjects having similar values of the propensity score share the same distribution of characteristics (covariates). Therefore, one can eliminate the treatment-selection bias in the observational study by controlling for the propensity score. In this chapter, we will evaluate the effectiveness of propensity score adjustment by matching and weighting based on propensity score estimates from a few different approaches, as detailed below.

Logistic regression (LR) is typically used to estimate propensity scores where treatment status is regressed on a set of observed covariates (Austin & Stuart, 2015). LR is a strong tool for statistical analysis; however, as McCaffrey et al. (2005) points out, large numbers of covariates tend to hurt its ability to accurately estimate propensity scores, as a result of multicollinearity. Non-linearities and interaction terms can also increase the number of covariates and can add to overfitting if iterative model-building and variable

DOI: 10.1201/9781003181187-22

selection are not performed, further affecting propensity score estimation. Due to the potentially abundant demographic and academic preparation variables in student success studies, these can be common hurdles in analysis. In order to account for large numbers of covariates, interactions, and non-linear terms, we look to the use of random forests (RFs). We have found great success in applying random forest in educational research and for analyzing student success from pedagogical interventions (Spoon et al., 2016; He et al., 2018).

Random forest is an ensemble of many decision trees and, as a non-parametric method, overcomes all the issues of LR that are mentioned above. RF has been recognized as an excellent predictive tool compared to other machine learning methods, see, for example, Fernandez-Delgado et al. (2014), and He et al. (2018). In addition, the superior performance of random forest can be achieved with little model tuning and/or calibration by the user, making it an ideal tool for education researchers when estimating the propensity score. In this chapter, we also propose an additional ensemble learning method that combines predictions from eight popular machine learning methods (Hastie et al., 2016; James et al., 2013), including logistic regression, random forest, boosting, bagging, *k*-nearest neighbor, support vector machines (SVM), neural network, and naive Bayes. A combination of these base learner predictions should provide a more accurate propensity score estimation than any one base learner such as random forest (for discussion, see Beemer et al., 2017).

The goal of this research is threefold. The first goal is to compare accuracy of the propensity score estimates from three approaches: logistic regression, random forest, and the ensemble learner (EL). The second goal is to compare precision of the treatment effect estimates based on propensity score estimates from these three approaches, in tandem with propensity score adjustments from the literature, including matching, weighting, variance stabilization, and truncation. The first two goals are achieved by a large-scale simulation study. The third goal is achieved by applying select methods to evaluate the effectiveness of an educational intervention from San Diego State University.

The chapter is organized as follows. In Section 18.2, we provide relevant information about propensity score, random forest, and the proposed ensemble learner. In Section 18.3, we present the design and results of the simulation study. In Section 18.4, we provide a student success case study of observational data. We conclude the chapter in Section 18.5 with a summary of results and some discussions.

18.2 Methods

18.2.1 Propensity Score

Propensity score, e_i, is the probability of a subject's assignment to a treatment, while taking into account the subject's characteristics:

$$e_i = P(Z_i = 1 \mid X_i). \tag{18.1}$$

Here Z_i is a binary treatment indicator, $Z_i = 1$ if a subject is in the treatment group and $Z_i = 0$ if a subject is in the control group, and X_i is a vector of all observed variables other than treatment assignment.

18.2.1.1 Assumptions for Propensity Score-based Methods

Propensity score based methods rely on four primary assumptions, as described below.

First, we assume that the treatment does not change in application across subjects. We assume no spillover effects, so a subject's treatment is not impacted by the treatment application on another subject. This assumption is standard in causal inference and named SUTVA – the stable unit treatment value assumption. In educational data mining applications, we must ensure interventions, instructional designs, intelligent tutoring systems, etc. follow a consistent template or rubric. We put particular attention on teacher training, if not common instructor, and intervention system version control to satisfy SUTVA. One also must be careful with student collaborations as that may lead to spillover effects.

Second, we assume that the outcome observed is the potential outcome that would be observed under the applied treatment. This assumption is labeled consistency and falls into counterfactual thinking for causal inference. A subject may receive any of the treatments under study. Each subject thus would realize a potential outcome for each of the treatment applications. In typical educational data mining practice of course, a subject is exposed to only one treatment. Consistency assumes that under the treatment actually received, a subject's observed outcome is the same as the subject's potential outcome. We find this assumption is always satisfied in studies of AI in education given careful definition of, protocols for, and application of the treatment regimes.

Third, we assume that every subject has a non-zero probability of receiving each treatment:

$$0 < P(Z = 1 \mid X) < 1. \tag{18.2}$$

This assumption is called positivity. In typical efficacy studies of AI in education, students have access or may receive any of the treatments under consideration, satisfying positivity.

Fourth, we assume that the treatment assignment is independent of the outcome conditional on the observed covariates:

$$\{Y(1), Y(0)\} \perp\!\!\!\perp Z \mid X, \tag{18.3}$$

where $Y(1)$ and $Y(0)$ are the possible subject outcomes for the treatment and control groups. This assumption is known as the "no unmeasured confounders" assumption, meaning all variables that affect the outcome and treatment assignment have been measured. Together, these latter two assumptions establish if the treatment assignment is strongly ignorable. Unfortunately, this fourth assumption cannot be verified in practice (Zhang et al., 2012). The usual tactic in causal inference is to collect and curate as many relevant inputs as possible, and run careful thought exercises on potential missing confounders.

For more details and mathematical buildup of these assumptions in observational studies, we refer the reader to Wilke et al. (2021) and the references therein. The bottom line for studies of AI in education is that with these assumptions, conditioning on the propensity score supports obtaining unbiased average treatment effect estimates (Rosenbaum & Rubin, 1983).

18.2.1.2 Propensity Score Matching

The propensity score matching method entails the following steps. First, the propensity score is estimated for each subject via a predictive model. Second, starting with a randomly selected treated subject, the subject is matched to a subject in the control group with the nearest propensity score; both the treated and control subjects are removed from the pool of future matches. This process is continued until all treated subjects are matched with a control subject. The final matched set will have an equal number of treated to control subjects, with the goal of having a balanced distribution for each covariate between the two treatment groups.

18.2.1.3 Inverse Probability of Treatment Weighting

Austin and Stuart (2015) review inverse probability of treatment weighting, variance stabilization, and truncation of weights as ways to better estimate treatment effects in observational studies. These methods give an educational researcher alternatives to propensity score matching, while still accounting for observed covariates in the study. This section will walk-through the three propensity score weighting methods.

Inverse probability of treatment weighting (IPTW) adjusts the underrepresented and overrepresented subjects within the control and treatment groups by assigning weights:

$$w_i = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i}, \tag{18.4}$$

where Z_i and e_i are the treatment indicator and propensity score ($i = 1, \ldots, n$, for n observations), respectively. IPTW gives higher weights to those in the treatment group with low propensity scores and those in the control group with high propensity scores, giving a more accurate estimation of the treatment effect (Rosenbaum, 1987). We propose to use IPTW in combination with regression models, as referenced in Austin and Stuart (2015), to improve estimation of causal treatment effects compared to propensity score matching.

When the propensity scores are close to zero in the treatment group and close to one in the control group, IPTW will assign a large weight to those subjects. Thus, a small group of subjects may carry a large proportion of the propensity score weight leading to potentially poor treatment effect estimation. To counter the increase in variability, we stabilize the weights by multiplying the treatment indicator, Z_i, by the marginal probability of treatment, $Pr(Z = 1)$, and multiplying the control indicator, $1 - Z_i$, by $Pr(Z = 0)$ (Robins et al., 2000). The adjusted weight is

$$w_i = \frac{Z_i Pr(Z = 1)}{e_i} + \frac{(1 - Z_i) Pr(Z = 0)}{1 - e_i}. \tag{18.5}$$

Lee et al. (2010) propose trimming or truncating the weights assigned by IPTW, again to prevent extreme weights from being assigned when the propensity scores are close to zero or one. The truncation is done by designating a minimum and maximum threshold, and if weights exceed the threshold, they are set to that threshold (Cole & Hernan, 2008; Lee et al., 2010).

18.2.2 Random Forest

Random forest uses bootstrap samples as training data to grow individual trees and then combines predictions from all trees by model averaging. Figure 18.1 presents an illustrative decision tree for presentation of the key terminology and concepts. This tree classifies students based on whether they went to a success program or not in an introductory statistics course. A tree consists of the root node (oval) and internal nodes (rectangles), each characterized by decision rules by which students are sent down the tree either to the left or to the right. For example, the root node splits students according to performance on a beginning

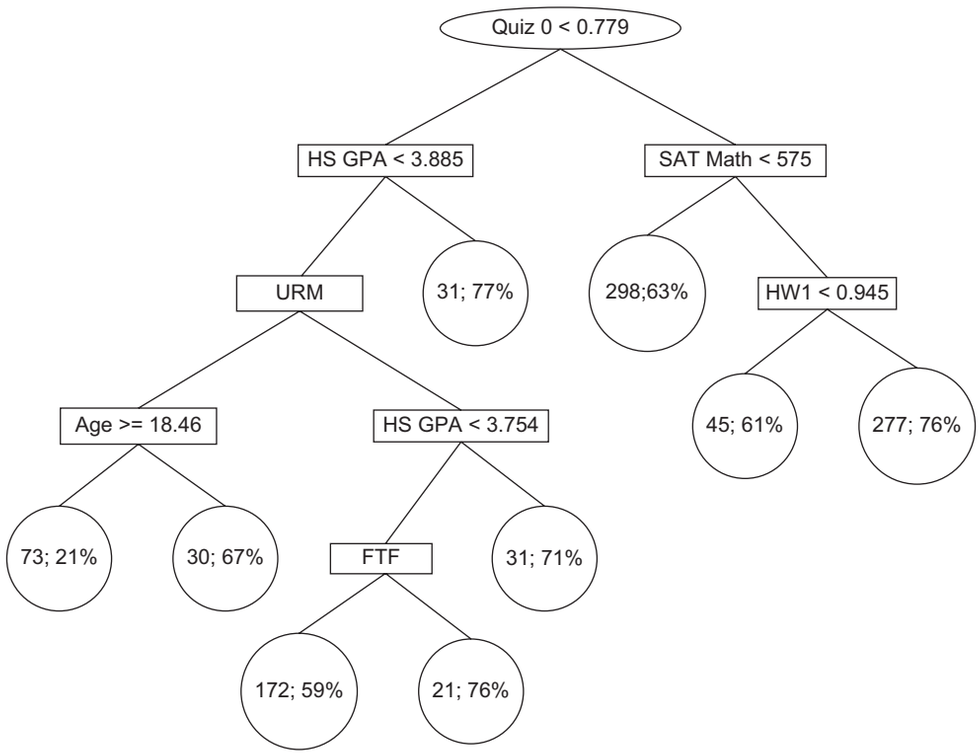

FIGURE 18.1
Illustrative decision tree to exposit the role of the root node (oval), internal nodes (rectangles), and terminal nodes (circles), and the decision rule determining splits and progression of observations down a tree. This graphic is of a classification tree with outcome (output) being treatment assignment in an introductory statistics course. The variables (inputs) on which the tree is grown are beginning of semester math readiness (Quiz 0), high school grade point average (GPA), SAT math score, underrepresented minority status (URM), student age, identification as a first-time freshman (FTF), and score on the first homework of the semester (HW1).

of semester math readiness assessment (Quiz 0). Students scoring below 0.779 are sent left down the tree, while students scoring at least 0.779 on the assessment are sent to the right down the tree. In this split, the root node is called the "parent," and the two internal nodes to which students may be sent are called the "children." The students continue to progress down the tree in this way until reaching a terminal node, the circles in Figure 18.1. For illustration purposes, in each terminal node, we show the number of students and the percentage of students attending the success program (treated).

At each split of a tree, a subset of covariates are chosen randomly. All the splitting rules based on these variables are assessed. A decision rule is chosen to split the data from this set of splitting rules according to a specified optimization criterion (see more on this below). Each split within a tree is binary and leads to another split or terminal node. The tree continues to grow until the terminal nodes are homogeneous, a preset tree depth is met, or a minimum number of observations is reached in a terminal node. For random forest, this process is repeated to grow many trees, thus creating a forest of decision trees. As an additional randomization element to construct a forest, each tree is grown on data randomly sampled with replacement from the original dataset. This bootstrap sample is the same size as the original dataset, mimicking data replication to allow for variety among the trees (decision rules) in the forest.

For each tree, the response variable (outcome/output) is typically predicted as a majority vote or average, for classification or regression respectively, based on the observations in each terminal node (Breiman, 2001). However, Malley et al. (2012) found that using the average outcome, as implemented in our study, outperformed the majority vote for classification problems. For the forest, the predicted values from each tree are averaged across all trees. Unlike logistic regression, random forest is unaffected by monotonic transformations of input variables, accommodates interactions among input variables through the recursive bisecting of the data, and reduces overfitting by taking bootstrap samples for growing each tree (Lee et al., 2010). A useful feature of random forest is the so-called out-of-bag (OOB) sample, which are the observations not included in the bootstrap sample and can

be used as a built-in validation sample to assess prediction accuracy.

For propensity score estimation, the response variable we intend to predict is the treatment assignment. The best cut point for each node is determined by maximizing the reduction in the within-node impurity between the parent node and children nodes. In this project, we utilize the Gini index–based splitting criterion to evaluate each candidate split (Breiman et al., 1984). For binary response variable, the Gini index of node t can be written as follows:

$$i(t) = 1 - \{p(1 \mid t)\}^2 - \{p(0 \mid t)\}^2, \tag{18.6}$$

where $p(0 \mid t)$ and $p(1 \mid t)$ are the proportion of treated and untreated subjects in the node t. Smaller values of $i(t)$ indicate purer nodes with respect to the response. The goodness-of-split criterion $\Delta i(s,t)$ is defined as follows:

$$\Delta i(s,t) = i(t) - \{p_L i(t_L) + p_R i(t_R)\}, \tag{18.7}$$

where t_L and t_R are the descendants of node t split by ss, and the weights, p_L and p_R, are the proportion of subjects in node t that are partitioned into the left child node t_L and right child node t_R, respectively. This splitting criterion shows the impurity difference between the parent node and two weighted child nodes, and the best split is the one having the largest $\Delta i(s,t)$.

After all trees are constructed, information provided by the terminal nodes is used to estimate the propensity score for each observation in the dataset. Observations ending up in the same terminal node have the same estimated propensity score, which is equal to the percentage of treated subjects in the node. The propensity score estimate for each subject based on the random forest is the average propensity score over the classification trees in the forest. In this chapter, we grow forests of 500 trees.

18.2.3 Ensemble Learning

Random forest in itself is an ensemble learner as we are making predictions across a set of trees in the forest. More generally, ensemble learning entails fitting multiple individual learners to the data, and combining predictions from these individual learners into an ensemble prediction. The idea of an ensemble learner is as follows. Using cross-validation and looping over the data, a predicted value is obtained for each observation based on each individual learner. Wolpert (1992), Breiman (1996), and LeBlanc and Tibshirani (1996) propose that the ensemble may combine the predictions from individual or base learners using a ridge regression, the so-called meta-learner. Ridge regression uses a penalty function to minimize coefficients of covariates that are weak predictors of the outcome (James et al., 2013, Chapter 6). In this case, the predictions from the individual learners are the covariates and the coefficients act as the weights, which become the ensemble learner via linear combination.

We propose a modified version of the ensemble learner presented in Beemer et al. (2017). The ensemble learning method starts with K-fold cross-validation: randomly split the data into K subsets of approximately equal sizes, removing one subset and training the base learners on the remaining subsets. The trained base learners are then used to make predictions for the subset that was removed. This "leave-one-out" method is repeated using the next subset until predictions are made for all observations. The predictions from base learners are then stacked, with predictions from each base learner forming a column. But instead of using a ridge regression to weigh the predictions as described above, a random forest is built to regress the true outcome against the predictions. Random forest has proven to be a reliable meta-learner in medical research (Wang et al., 2019), and a good alternative to regression methods in educational research (Spoon et al., 2016; He et al., 2018). We expect this modified ensemble learner to achieve strong predictive performance.

Algorithm 1: Ensemble Learner

1. Identify L base learners.
2. Randomly partition data into k subsets of approximately equal sizes.
3. **for** $k = 1, \ldots, K$ **do**
4. Leave the kth subset out as test set, remaining subsets are the training set.
5. **for** $j = 1, \ldots, L$ **do**
6. Train the j^{th} base learner on training set.
7. Predict test set using trained base learner.
8. Stack predictions from each base learner as a column in a data frame.
9. Build a random forest using the stacked predictions from the base learners as inputs.

The algorithm for the ensemble learner is presented in Algorithm 1. The ensemble will have the ability to combine the predictions from L machine learning methods. In our software, we provide the user with eight base learners: logistic regression, random forest, boosting, bagging, k-nearest neighbor, support vector machines, neural network, and naive Bayes.

All codes performed in the project are developed in the statistical freeware R. Random forest is constructed through the R package PartyKit with modifications. The R package MatchIt is used to obtain the samples matched by propensity scores estimated using LR, RF, and our custom-made ensemble learner. The R software we wrote for the ensemble learner is publicly available at our GitHub depository (Beemer, 2021).

18.3 Simulation Study

In this section we present a simulation study to compare the propensity score estimates from logistic regression, random forest, and the ensemble learner. The ultimate goal is to evaluate the precision in treatment effect estimates when using different modeling approaches for PS estimation, combined with various propensity score adjustment methods. We start with data generation for the observational study.

18.3.1 Data Generation

We generate data in the following order.

18.3.1.1 Generating Covariates (X)

Eight covariates (X_1–X_8) were generated independent of specific probability distributions. Variable X_1 was generated as a binary variable from a Bernoulli distribution with probability of success at $p = 0.5$. Variable X_2 was generated as a nominal variable with five categories (A, B, C, D, E), with each category at different likelihoods to occur, (10%, 20%, 30%, 20%, 20%). Variables X_3 and X_4 were generated independent of a discrete uniform distribution from 0 to 1 with increments of 0.2, and treated as ordinal variables with five levels (0.2, 0.4, 0.6, 0.8, 1.0). The last four covariates (X_5 –X_8) were designed to mimic continuous variables and were simulated by discrete uniform distributions from 0 to 1 with increments of 0.02.

18.3.1.2 Generating Treatment Assignment Indicator (Z)

Following Setoguchi et al. (2008), the true propensity score, or the probability of treatment assignment given covariates, is assumed to follow the logistic regression model:

$$P(Z=1\mid X)=\frac{1}{1+e^{-\beta f(X)}}. \tag{18.8}$$

Setoguchi et al. (2008) uses seven models in their study, in which the function $f(\cdot)$ has varying degrees of additivity and linearity, with non-additivity and non-linearity comprising two-way interactions and quadratic terms. In this chapter, we consider four models from Setoguchi et al. (2008) and add one additional model (model E) which contains three-way interactions and non-linearity terms other than the quadratic form. The five treatment-selection models are given as follows:

A. *Additivity and linearity (main effects only)*

$$P(Z=1\mid X)=(1+\exp[-(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8)])^{-1}. \tag{18.9}$$

B. *Moderate non-linearity (three quadratic terms)*

$$P(Z=1\mid X)=(1+\exp[-(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_9X_3^2+\beta_{10}X_5^2+\beta_{11}X_7^2)])^{-1}. \tag{18.10}$$

C. *Mild non-additivity (four two-way interaction terms)*

$$P(Z=1\mid X)=(1+\exp[-(\beta_0+\beta1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_9X_3X_4+\beta_{10}X_4X_5+\beta_{11}X_5X_6+\beta_{12}X_6X_7)])^{-1}. \tag{18.11}$$

D. *Moderate non-additivity and non-linearity (ten two-way interaction terms and three quadratic terms)*

$$P(Z=1\mid X)=(1+\exp[-(\beta_0+\beta1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_9X_3^2+\beta_{10}X_5^2+\beta_{11}X_7^2+\beta_{12}X_3X_4+\beta_{13}X_4X_5+\beta_{14}X_5X_6+\beta_{15}X_6X_7+\beta_{16}X_7X_8+\beta_{17}X_3X_8+\beta_{18}X_5X_7+\beta_{19}X_4X_8+\beta_{20}X_3X_5+\beta_{21}X_6X_8)])^{-1}. \tag{18.12}$$

E. *Severe non-additivity and non-linearity (six two-way interaction terms and four three-way interaction terms; one quadratic term, one cubic polynomial, and one square root term)*

$$P(Z=1\mid X)=(1+\exp[-(\beta_0+\beta 1X_1+\beta_2X_2+\beta_3X_3 +\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8 +\beta_9X_4^2+\beta_{10}X_6^2+\beta_{11}\sqrt{X_8}+\beta_{12}X_3X_4 +\beta_{13}X_4X_5+\beta_{14}X_5X_6+\beta_{15}X_6X_7 +\beta_{16}X_7X_8+\beta_{17}X_3X_8+\beta_{18}X_3X_5X_7 +\beta_{19}X_4X_6X_8+\beta_{20}X_3X_4X_5+\beta_{21}X_6X_7X_8)])^{-1}. \tag{18.13}$$

The true propensity score was used as the parameter, *p*, in a Bernoulli distribution to generate the treatment assignment for each subject in the dataset. In our experience with educational interventions, the proportion of students in the treatment group is typically within the 20–30% range. The coefficients (shown in Table 18.1) in the treatment-selection models were chosen so that the probability of being allocated to treatment was about 25%.

18.3.1.3 Generating Outcome (Y)

Based on covariates (X_1-X_8) and treatment assignment indicator (Z), the continuous outcome variable, Y, was generated as follows:

Model 1

$$Y=\alpha_{00}+\alpha_0Z+\alpha_1X_1+\alpha_2X_2+\alpha_3X_3+\alpha_4X_4+\alpha_5X_5 +\alpha_6X_6+\alpha_7X_7+\alpha_8X_8+\varepsilon. \tag{18.14}$$

Model 2

$$Y=\alpha_{00}+\alpha_0Z+\alpha_1X_1+\alpha_2X_2+\alpha_3X_3^2+\alpha_4X_4^2+\alpha_5\ln X_5 +\alpha_6\sqrt{X_6}+\alpha_7X_7X_8+\alpha_8X_3X_7+\varepsilon, \tag{18.15}$$

where $\varepsilon\sim N(0,1)$. In Model 1, a simple linear association is assumed between predictors and the outcome, while Model 2 involves several non-linear terms and two-way interactions, in order to examine the performance of different methods for treatment effect estimation in a more complex data structure. The true treatment effect (α_0) was fixed at 1.5 for both outcome models. The other coefficients in the models($\alpha_1-\alpha_8,\alpha_{00}$) were set to 0.5, 0.3, 0.7, 0.6, 0.1, –1.2, –0.5, –1, and 0.5, respectively.

With five models (A–E) for treatment assignment, combined with two models (1–2) for the outcome, a total of ten models were used in our simulation study. These models are denoted as models A–E and A¢–E¢ hereafter. For example, model A assumes perfect additivity and linearity for both the treatment assignment model and the outcome model, while model E¢ assumes most severe non-additivity and non-linearity for both the treatment assignment and outcome models. For each simulation scenario, 100 datasets of size $n=500$ were simulated.

18.3.2 Simulation Study Results

Propensity scores were predicted from logistic regression, random forest, and the ensemble learner for the true propensity score models A–E, and from these predictions a mean squared error (MSE) was computed. Table 18.2 shows that the ensemble learner performs the best, and random forest performs the second best, for every treatment assignment model. Logistic regression underperforms even for Model A, which is a logistic regression model with perfect additivity and linearity. The excellent performance of the ensemble learner, especially for more complex models, supports the idea of using an ensemble learner over logistic regression for propensity score matching and weighting techniques.

Table 18.3 presents bias and mean squared error (MSE) for the estimated treatment effect using

TABLE 18.1

Coefficients Used in the Data Generation Models

Model	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
A	–0.5	–0.5	1.2	–1.0	–0.62	–0.7	–0.4	0.6	0.2	•	•
B	–0.5	–0.5	–1.2	–1.2	–0.72	0.7	0.4	0.6	0.2	0.3	–0.4
C	–0.5	–0.5	–1.2	–1.2	–0.72	0.7	0.4	0.6	0.2	0.3	–0.4
D	–0.5	–0.5	–1.2	–1.2	–0.62	0.7	0.4	0.6	0.2	0.3	–0.4
E	–0.5	–0.5	–1.2	–1.2	–0.72	0.7	0.4	0.6	0.2	0.3	–0.4
Model	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}	β_{18}	β_{19}	β_{20}	β_{21}
A	•	•	•	•	•	•	•	•	•	•	•
B	1.1	•	•	•	•	•	•	•	•	•	•
C	1.1	0.46	•	•	•	•	•	•	•	•	•
D	1.1	–0.2	0.42	–0.8	0.9	–1	0.32	–0.45	0.36	–0.47	0.35
E	1.1	–0.2	0.42	–0.8	0.9	–1	–0.32	–0.45	–0.36	0.47	0.35

TABLE 18.2

Mean Squared Error (MSE) for the Propensity Score Estimates by Logistic Regression, Random Forest, and Ensemble Learner for Models A–E ($n = 500$)

	A	B	C	D	E
LR	0.041	0.052	0.054	0.048	0.047
RF	0.039	0.048	0.049	0.044	0.044
EL	0.018	0.017	0.017	0.017	0.018

propensity score matching and weighting, including four different weighting schemes (inverse probability of treatment weighting, variance stabilization of weights, weight truncation, variance stabilization with truncation), as detailed in Section 18.2.1. The methods for propensity score estimation include logistic regression, random forest, and ensemble learner. These results offer comparisons among models (logistic regression, random forest, and ensemble learning) and between matching and weighting.

It can be seen from Table 18.3 that the best performance, as signified by the smallest MSE, is almost always by the ensemble learner, especially when the ensemble learner is combined with propensity score weighting using variance stabilization and/or truncation of weights. Random forest, combined with propensity score weighting, also performs well. Comparing propensity score matching and weighting, weighting appears to be the clear winner, especially when using variance stabilization and/or truncation of weights.

18.4 Students' Success Case Study

In a "State of the CSU" address, Dr. Timothy White, the former Chancellor of the California State University system, states "The California State University is key to California's brightest and most hopeful future, opening the door to educational opportunities for all and transforming the lives of students and their families." With this mandate, San Diego State University offers many student success interventions such as supplemental instruction (see Guarcello et al., 2017, and the references therein), which is offered to students currently enrolled, or even pre-enrollment interventions that aim to help students to attend SDSU. We look at one such student success intervention that is offered to students who are from underprivileged communities and are given a path that optimally could afford them the opportunity for entrance into higher education.

Table 18.4 presents a comparison of students supported by the student success intervention and their peers not assisted by the intervention. This snapshot of a few student background characteristics shows that those in the intervention have a higher rate of being first-generation college students (i.e., first in their immediate family to go to college) and underrepresented minorities (an ethnicity categorization defined by the California State University System). They have lower mean SAT scores, slightly lower mean high school grade point average (GPA), and tend to earn, and transfer, fewer college-level course units than their peers at the university.

Table 18.5 investigates the balance in the demographic and background variables before and after matching, between treated and control groups. Before propensity score matching, the standardized mean difference (SMD), defined as the difference between the two sample means divided by the pooled standard deviation, is large for all but two variables in Table 18.5, using a value of SMD below 20 for balanced samples (Austin, 2009; Hillis et al., 2021). After propensity score matching, the standardized mean difference decreases for all covariates. Matching in this study does a very good job of balancing the background characteristics between the treated and control groups, with all SMD values below 20 after matching.

To evaluate the success of the intervention, we examine the effect of the intervention on the student GPA at the end of their second semester at the university, specifically the GPA for courses taken on campus. Table 18.6 presents the estimated treatment effect of the intervention, and the associated p-value and 95% confidence interval. Based on the results using propensity score weighting with variance stabilization and truncation, those students who participated in the student success intervention had on average an increase of 0.053 (with a 95% confidence interval of –0.024 to 0.130) in their end of second semester GPA compared to those students who did not participate in the intervention, accounting for all other possible factors. The results based on propensity score matching are similar to those from weighting, and the ensemble learner was used for propensity score estimation.

We note that the treatment effect confidence intervals cover zero, suggesting a failed student success intervention. However, since students provided with the intervention are perceived to be at a distinct academic disadvantage due to their socioeconomic background, an "on par" result is a success. These results show that by participating in the student success intervention, program students were able to match their peers in GPA at the end of their second semester.

18.5 Discussion

Evaluations and assessments of student success interventions often require observational studies.

TABLE 18.3

Bias and Mean Squared Error (MSE) for the Treatment Effect Estimates by Logistic Regression, Random Forest, and Ensemble Learner Using Various Propensity Score Matching and Weighting Techniques for All Models A–E and A′–E′ ($n = 500$)

		Matching			IPTW			Variance Stabilization			Weight Truncation			Variance Stabilization with Truncation		
		LR	**RF**	**EL**	**LR**	**RF**	**EL**	**LR**	**RF**	**EL**	**LR**	**RF**	**EL**	**LR**	**RF**	**EL**
A	BIAS	–0.016	–0.021	0.005	–0.028	–0.017	–0.004	–0.027	–0.015	0.003	–0.030	–0.025	–0.013	–0.029	–0.022	–0.007
	MSE	0.028	0.035	0.020	0.022	0.027	0.010	0.023	0.027	0.009	0.020	0.021	**0.008**	0.020	0.026	0.010
B	BIAS	0.012	0.029	0.044	0.020	0.025	–0.025	0.021	0.027	–0.004	0.009	0.015	0.007	0.013	0.026	0.019
	MSE	0.043	0.024	0.027	0.045	0.042	0.051	0.047	0.041	0.042	0.036	0.035	0.023	0.037	0.038	**0.020**
C	BIAS	–0.049	–0.037	–0.060	–0.041	–0.037	0.002	–0.040	–0.04	0.006	–0.055	–0.052	–0.039	–0.045	–0.048	–0.043
	MSE	0.033	0.057	0.023	0.026	0.026	0.011	0.026	0.028	**0.010**	0.029	0.030	**0.010**	0.029	0.028	0.012
D	BIAS	–0.005	0.001	0.028	–0.007	–0.005	–0.020	–0.004	0.001	0.001	–0.005	–0.008	–0.002	–0.009	–0.002	0.010
	MSE	**0.016**	0.029	**0.016**	0.045	0.034	0.040	0.047	0.031	0.032	0.028	0.028	0.022	0.028	0.027	0.017
E	BIAS	–0.029	–0.049	–0.020	–0.011	–0.011	–0.016	–0.009	–0.016	–0.033	–0.012	–0.016	–0.015	–0.011	–0.017	–0.031
	MSE	0.029	0.015	0.013	0.010	0.009	0.016	0.010	0.010	0.015	0.010	0.010	**0.007**	0.010	0.010	**0.007**
A′	BIAS	0.063	0.049	0.054	0.098	0.075	0.101	0.098	0.069	0.094	0.064	0.068	0.067	0.068	0.065	0.058
	MSE	0.029	0.043	0.019	0.039	0.035	0.032	0.040	0.034	0.030	0.028	0.030	0.017	0.031	0.033	**0.016**
B′	BIAS	–0.078	–0.093	–0.074	–0.102	–0.095	–0.033	–0.104	–0.092	–0.016	–0.101	–0.095	–0.089	–0.088	–0.085	–0.079
	MSE	0.025	0.019	0.010	0.026	0.018	0.007	0.026	0.015	**0.006**	0.021	0.018	0.014	0.016	0.014	0.011
C′	BIAS	–0.044	–0.043	–0.044	0.042	0.009	0.031	0.047	0.001	0.019	0.004	0.005	0.021	0.006	–0.003	0.012
	MSE	0.015	0.014	0.019	0.012	0.011	0.010	0.013	0.012	0.011	0.010	0.011	**0.006**	0.012	0.011	0.007
D′	BIAS	0.007	0.014	0.025	0.021	0.011	–0.033	0.022	0.013	–0.041	0.014	0.014	0.019	0.010	0.016	0.024
	MSE	0.012	0.016	0.008	0.019	0.014	0.040	0.020	0.013	0.042	0.013	0.012	0.008	0.012	0.011	**0.007**
E′	BIAS	0.055	0.048	0.043	0.034	0.035	0.080	0.033	0.038	0.082	0.034	0.039	0.036	0.038	0.037	0.041
	MSE	0.009	0.015	0.006	0.009	0.011	0.017	0.008	0.010	0.018	0.009	0.010	**0.005**	0.008	0.009	**0.005**

TABLE 18.4

Summary of Student Characteristics for Treatment and Control Groups

	Control	Treated
First generation	15.8%	21.0%
Underrepresented minority	32.4%	62.9%
SAT score	1209.1 (156.3)	1145.3 (85.1)
High school GPA	3.7 (0.3)	3.5 (0.3)
Transfer units	22.3 (26.3)	10.8 (11.2)

Mean and (standard deviation) reported for continuous variables, and percentage reported for categorical variables.

TABLE 18.5

Standardized Mean Difference before and after Matching

Covariate	Before	After
Age	62.8	6.7
Gender	10.4	3.2
SAT score	50.8	7.3
High school GPA	32.8	18.4
Incoming units	57.1	8.3
First Generation	13.3	2.6
Underrepresented minority	64.1	7.7
Hispanic	72.2	15.2

Propensity score–based adjustments are powerful tools to derive unbiased estimates of the treatment effect from observational studies. In this chapter, we review two methods for propensity score estimation, including logistic regression and random forest, and propose our own custom-made ensemble learner combining prediction results from eight popular machine learning methods. In addition, we discuss propensity score matching and inverse probability of treatment weighting, including variance stabilization and truncation of weights, as means to improve performance of propensity score weighting.

A large-scale simulation study is conducted to compare the three modeling approaches for propensity score estimation: LR, RF, and EL; and to compare performance of propensity score matching and weighting in conjunction with the modeling approaches. The simulation results show that the ensemble learner provides the most accurate estimates of the propensity score under all model configurations, followed by random forest as the second best performer. In terms of accuracy of treatment effect estimation, the ensemble learner combined with propensity score weighting, incorporating variance stabilization and truncation, is an overall top performer. Random forest combined with propensity score weighting also performs reasonably well. Between propensity score matching and weighting, we recommend propensity score weighting using variable stabilization and truncation of weights.

The ensemble learner based propensity score matching and weighting methods are applied to a student success intervention at San Diego State University for underserved students before enrollment at SDSU. The ensemble learner–based propensity score matching is able to largely eliminate the imbalance in student background variables between students in the intervention and their peers not in the intervention. A study of the effect of intervention, using propensity score matching and weighting, shows that the intervention successfully removes academic disadvantage among participates, a consequence of their lower socioeconomic status and less prepared academic background, so that the participants are able to perform as well as the general student population at SDSU by the end of their second semester at the university.

In terms of performance comparisons of existing machine learning methods, Fernandez-Delgado et al. (2014) performed a broader study of 179 classifiers from 17 machine learning families and result found that random forest performed the best overall, followed by boosted trees, neural network, and SVM. In comparison, naive Bayes, logistic regression, and decision tree do not perform as well. Considering propensity score weighting specifically, Lee et al. (2010) conducted a well-cited simulation study comparing the performance of logistic regression, decision tree, bagged and boosted trees, and random forest. Their recommendation was to use boosted trees and random forest for their consistent superior performance. More recently, Cannas and Arpino (2019) extended the simulation study by Lee et al. to include both propensity score matching and weighting, while adding two new machine learning methods for comparison: neural network and naive Bayes. They also found random

TABLE 18.6

Student Success Treatment Effect, *p*-Value, and 95% Confidence Interval from Matching and Variance Stabilization with Truncation

Method	Treatment Effect	*p*-Value	95% Confidence Interval
Matching	0.063	0.196	−0.033, 0.158
Variance stabilization with truncation	0.053	0.179	−0.024, 0.130

forest to have the best overall and most consistent performance, followed by neural network and logistic regression. In summary, random forest, boosted trees, and neural network all seem to perform well in general, with boosted trees and neural network requiring more user input and calibration. For the less experienced user of machine learning methods, without having to code their own ensemble or performing extensive calibration, we recommend random forest for propensity score estimation for its superior predictive power and relative ease of implementation.

We used a sample size of 500 with eight features in simulations presented in this chapter. For the simulation studies presented in Lee et al. (2010) and Cannas and Arpino (2019), sample sizes of 500, 1,000, and 2,000 were considered with ten features. These sample sizes and number of features were selected as they were similar to the observational study data under consideration and the recommended machine learning methods do pretty well under these configurations. Lee et al. point out that as the sample size increases, the comparative performance of the machine learning algorithms did not change, while the accuracy of treatment effect estimates improved for all methods. This generally agrees with our own experiences, see, for example, Autenrieth et al. (2021), in which an in-depth simulation study was provided. In addition, machine learning methods can generally perform well over a wide range of sample sizes and feature space.

It is important to point out that no method performs the best for all situations. Traditional methods such as logistic regression perform well when the models can be correctly specified, while machine learning methods such as random forest and ensemble learner have a distinct advantage with complex data structure since they are non-parametric in nature and hence more flexible. This trend can be seen from the simulation results presented in Table 18.3. Since one does not know the true model formats in the real world, we recommend that different methods should be evaluated for the specific dataset at hand with performance judged based on cross-validation or a test sample.

Quantification of uncertainty with machine learning methods is more complex and much less routinely performed compared to classical regression methods. However, there have been some recent advances on statistical inferences for random forests, interested readers are referred to Mentch and Hooker (2016), Wager and Athey (2018), Athey et al. (2019), Lu and Hardin (2021), and references therein. In the context of propensity score weighting and matching, machine learning methods are used only to obtain more accurate estimates of the propensity score. Since the ultimate goal is to reduce or eliminate the selection bias from the observational study, it is crucial that the predicted propensity scores can help achieve well-balanced covariates between treatment groups.

Acknowledgment

This research was supported in part by the National Science Foundation grant 1633130.

References

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148–1178.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679.

Autenrieth, M., Levine, R. A., Fan, J. J., & Guarcello, M. A. (2021). Stacked ensemble learning for propensity score methods in observational studies. *Journal of Educational Data Mining*, *13*, 24–189.

Beemer, J. (2021). Ensemble learner codes: GitHub repository. https://github.com/jbeemer05/match_ED

Beemer, J., Spoon, K., He, L., Fan, J., & Levine, R. A. (2017). Ensemble learning for estimating individualized treatment effects in student success studies. *International Journal of Artificial Intelligence in Education*, *28*(3), 315–335.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*(1), 49–64.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olsen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth International Group.

Cannas, M., & Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal*, *61*(4), 1049–1072.

Cole, S. R., & Hernan, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*(6), 656–664.

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181.

Guarcello, M. A., Levine, R. A., Beemer, J., Frazee, J. P., Laumakis, M. A., & Schellenberg, S. A. (2017). Balancing student success: Assessing supplemental instruction through coarsened exact matching. *Technology, Knowledge and Learning*, *22*(3), 335–352.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

He, L., Levine, R. A., Fan, J. J., Beemer, J., & Stronach, J. (2018). Random forest as a predictive analytics alternative to regression in institutional research. *Practical Assessment, Research and Evaluation*, *23*(1), 1–16.

Hillis, T., Guarcello, M. A., Levine, R. A., & Fan, J. J. (2021). Causal inference in the presence of missing data using a random forest based matching algorithm. *Stat*, *10*(1), e326.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

LeBlanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, *91*(436), 1641–1650.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337–346.

Lu, B., & Hardin, J. (2021). A unified framework for random forest prediction error estimation. *Journal of Machine Learning Research*, *22*, 1–41.

Malley, J., Kruppa, J., Dasgupta, A., Malley, K., & Ziegler, A. (2012). Probability machines: Consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine*, *51*(1), 74–81.

Mccaffrey, D., Ridgeway, G., & Morral, A. (2005). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425.

Mentch, L., & Hooker, G. (2016). Quantifying uncertainty in random forest via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, *17*, 1–41.

Robins, J. M., Hernan, M., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(387), 394.

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

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety*, *17*(6), 546–555.

Spoon, K., Beemer, J., Whitmer, J. C., Fan, J. J., Frazee, J. P., Stronach, J., Bohonak, A. J., & Levine, R. A. (2016). Random forests for evaluating pedagogy and informing personalized learning. *Journal of Educational Data Mining*, *8*(2), 20–50.

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

Wang, Y., Wang, D., Geng, N., Wang, Y., Yin, Y., & Jin, Y. (2019). Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection. *Applied Soft Computing*, *77*, 188–204.

Wilke, M. C., Levine, R. A., Guarcello, M. A., & Fan, J. (2021). Estimating the optimal treatment regime for student success programs. *Behaviormetrika*, *48*(2), 309–343.

Wolpert, D. (1992). Stacked generalization. *Neural Networks*, *2*(2), 241–259.

Zhang, B., Tsiatis, A. A., Laber, E. B., & Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, *68*(4), 1010–1018.