ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2022-94304

A Comparison of Ensemble Learning Algorithms for Matching Weights Method: A Simulation Study

Eşleştirilmiş Ağırlıklandırılma Yöntem İçin Topluluk Öğrenme Algoritmalarının Karşılaştırması: Bir Simülasyon Çalışması

Hülya KOÇYİĞİT[a]

[a]Department of Mathematics and Science Education, Karamanoğlu Mehmetbey University Faculty of Education, Karaman, Türkiye

ABSTRACT Objective: This paper conducts thorough simulation research to assess the effectiveness of ensemble learning techniques and logistics regression models for estimating propensity score values used at the matching weighting under different propensity score model scenarios and various treatment scenarios considered. **Material and Methods:** This study underlines the significance and challenges of frequently disregarded overlap assumption. Offered method also is examined and focuses on the difficulties that non-overlap entails for inference. Monte Carlo simulations are used to generate data sets to analyze the causal effect of meeting in order that illustrates alternative strategies and pertaining aspects when highlighting positivity violations. **Results:** Here simulation results are illustrated to compare matching weight method under various machine learning methods in terms of root mean squared error (RMSE), SE of the treatment effects, and bias. Some ensemble learning algorithms for estimating propensity score (PS) values have rigorously outperformed than using the logistics regression method with or without existing a violation of the positivity the assumption under the different estimation PS models and various treatment models. The most complex treatment scenario tends to produce better results as measured by the SE, RMSE and bias than the less complex treatment scenarios. **Conclusion:** The findings summarize the conditions under which one technique may be anticipated to perform better than others without generalizing whether a method is always preferable to the other.

Keywords: Matching weighting; observational studies; ensemble learning models; propensity score; Monte Carlo simulation

ÖZET Amaç: Bu makalede dikkate alınan farklı eğilim puanı modeli senaryoları ve çeşitli tedavi senaryoları altında, eşleştirme ağırlıklandırılması metodunun kullanılan tahmini eğilim puan değerleri hesaplanması için topluluk öğrenme teknikleri ve lojistik regresyon modelinin etkinliğinin değerlendirmek için kapsamlı bir simülasyon çalışması yürütmektedir. **Gereç ve Yöntemler:** Bu çalışma, sıklıkla ihmal edilen örtüşme varsayımının önemini ve zorlukları vurgulamaktadır. Önerilen yöntemin de değerlendirilmesi ve nedensel çıkarımlarda örtüşmeme durumunun getirdiği zorluklara odaklanmıştır. Pozitiflik varsayımını vurgulanmasında alternatif stratejileri ve ilgili yönleri tanımlamak için nedensel etkiyi analiz etmek için Monte Carlo simülasyonundan elde edilen veri setleri kullanılır. **Bulgular:** Buradaki simülasyon çalışmasının sonuçları, farklı makine öğrenimi yöntemleri altında eşleştirme ağırlıklandırılması metodunun kök ortalama kare hatası [root mean squared error (RMSE)], tedavi etkilerinin SE ve göreceli ön yargı ölçülerine dayalı karşılaştırma yapmak için gösterilmektedir. Farklı tahmin eğilim skoru [propensity score (PS)] modelleri ve birçok tedavi modelleri altından pozitiflik varsayımının ihlali olsun veya olmasın PS değerlerinin tahmin etmek için kullanılan bazı topluluk öğrenme algoritmalarının, lojistik regresyon yönteminin kullanılmasından kesinlikle daha iyi performans göstermiştir. En karmaşık tedavi senaryosu, SE, RMSE ve yanlılıkla ölçüleri açısından daha az karmaşık tedavi senaryolarına göre daha iyi sonuçlar üretme eğilimindedir. **Sonuç:** Bulgular kısmıyla bir yöntemin diğerinden daha iyidir genelleştirmesini yapmadan, bir tekniğin diğerlerinden daha iyi bir performans göstermesinin beklenebileceğinin şartlarıyla özetlenmektedir.

Anahtar kelimeler: Eşleştirme ağırlıklandırılması; gözlemsel çalışmalar; topluluk öğrenme modeller; eğilim skoru; Monte Carlo simülasyon

Correspondence: Hülya KOÇYİĞİT
Department of Mathematics and Science Education, Karamanoğlu Mehmetbey University Faculty of Education, Karaman, Türkiye
E-mail: hk20902@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 12 Nov 2022 ***Received in revised form:*** 03 Feb 2023 ***Accepted:*** 06 Feb 2023 ***Available online:*** 07 Mar 2023

2146-8877 / Copyright © 2023 by Türkiye Klinikleri. This in an open
access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Observational research has contributed significantly to fields including public health, health economics, and medical science because of the unethical process and restricted access to databases in randomize control trials (RCTs). In order to decrease or eliminate the effects of confounding resulting from observed baseline factors in observational studies, applied researchers in the medical field are widely employing approaches based on the propensity score (PS). The PS estimates the probability that a treatment will be assigned based on observed baseline variables. Different PS based on the approaches has been studied to estimate causal inference in observational studies, such as matching, subclassification, inverse probability of treatment weighting (IPTW), and covariate adjustment. IPTW produces a reliable estimator of the average treatment effect (ATE) with the conditions of stable unit treatment value assumption (SUTVA), unconfoundedness, and overlap when PS models are determined correctly.[1-11] However, when the positivity assumption is violated, it can lead to incredibly high weights. In other words, a lack of overlap assumption happens when some individuals get treatment T_i, i=1,..,n when PS values $e(x_i)$ are close to 0 and 1. Unfortunately, the existing violation of the overlap assumption indicates that IPTW estimators can be excessively affected by a small number of extremely weighted observations, producing findings that are both biased and unstable. Researchers might not desire large weights because they would make determining the causes of events very difficult. Imbens and Rubin in 2015 state that this violation can happen for several reasons, including data limitations, a small sample size, incorrect PS model parameters, and incorrectly described relationships between the treatment/outcome and covariates.[12] Some papers-present a summary of traditional PS methods that have been put forth in the literature for assessing causal effects in the presence of overlap assumption.[13-15] Crump et al., Stürmer et al., Walker et al. have proposed different trimming methods, which are frequently employed to remedy positivity violations.[16-18] Trimming method is the process of locating a subset of individuals that seems to fail the positivity assumption, eliminating them from the data set, and making inferences about the remaining population. However, the employment of trimming techniques to deal with positive violations raises several possible challenges. Firstly, excluding individuals who commit positive violations reduces the sample size, which raises the possibility of the impact estimate's variance rising. Additionally, there is a strong correlation between sample size and how frequently positive violations occur by random. As the sample size decreases, new practical positive violations may be introduced, relying on how trimming is carried out. Furthermore, limiting the sample might have a causal impact on a population of restricted interest. Lastly, it can be challenging to interpret the parameter predicted when the criteria used to limit the sample includes a summary of high dimensional variables.[12] Despite the fact that trimming or traditional PS techniques may be suitable for structural violations, they are insufficient for actual violations of the positivity assumption. Researchers are prompted by this conflict to consider alternative intended samples for whom exposure impact might be more meaningfully and precisely explored in terms of bias, root mean squared error (RMSE), variance, or other measured metrics.

Later, Li and Greene proposed Matching Weights (MW) have been presented as alternatives to IPTW for overcoming overlap concerns.[19] They present the performance of pair-matching, inverse probability weight, double robust estimation, and matching weighting method that offers data analysts a new framework for determining whether the propensity model relies on the logistic regression is properly stated under the various simulation scenarios. There are some remarkable papers that employ the performance of MW and other balance weight approaches in literature after it is proposed by.[19] For example, MW method (with or without other balance weights methods) using logistic regression for an estimate of its PS values has been employed to compare the effectiveness of the bootstrap and asymptotic variance estimators, examined the performance of methods under the presence of overlap assumption, present extensive simulation under the violation of overlap assumption when having misspecified PS models scenarios.[20-22]

The balancing weights are a generic class of PS weights, many of which could be employed for covariate balance in observational studies, as I show out in this article. In spite of the potential benefits of MW, little is revealed regarding the comparative performance of the MW method in the literature. In the research area of new approach MW technique in the past decade, generating PS values are examined rely on the logistics regression, super learner or Gradient Boosting Machine (GBM) algorithms using simulated data and reported that super learner consistently gave the least biased results for implemented PSs on the MW method.[19-24] However, those researchers, who make investigations on MWs methods, do not consider ensemble learning techniques that thrive in classification and prediction processes and also, acquire favorable statistical properties. It is crucial to evaluate whether the efficiency of ensemble learning methods in PS estimation varies based on the data used. To fill this gap, the present work focuses on the ensemble learning techniques' ability to accurately estimate the causal effects of multi-component interventions. MW technique is then followed to decrease or eliminate bias between treatment groups. Finally, the performance of different methodologies is then compared, and I draw conclusions on how to evaluate the clinical efficacy of treatments.

MATERIAL AND METHODS

CAUSAL NOTATIONS

The Rubin Causal Model (RCM), which presents the causal inference framework relying on the series of bibliography, was first used.[1,25,26] Three main components are used to characterize the causal effects in RCM: assumed $\mathrm{i} = 1, \ldots, \mathrm{n}$ observations $(\mathrm{T_i}, \mathrm{X_i}, \mathrm{Y_i})$. This study considers a scenario in which there are n individuals, each of whom is indexed by $\mathrm{i} = 1, \ldots, \mathrm{n}$. In the binary case scenario, let T represents the observed treatment: T=1 for treatment group and T=0 for the control group and X is defined as a vector of observed variables. Rosenbaum and Rubin suggest two fundamental assumptions that enable proper causal inference.[1] The first assumption is the SUTVA which indicates the observed outcome as $\mathrm{Y_i} = \mathrm{Y_i}(1)\mathrm{T_i} + \mathrm{Y_i}(0)\mathrm{T_i}$. The second assumption is a strong ignorable treatment assignment also known as exchangeability and then, this assumption is expressed as follows in mathematical notation: $\mathrm{Y_i}(1), \mathrm{Y_i}(0) \perp \mathrm{T_i}|\mathrm{X_i}$. The exchangeability assumption assumes that exposure and outcomes, provided covariates, are independent. In other words, this assumption claims that all confounders are measured. Lastly, the overlap is called as the positivity assumption: $0 < P(\mathrm{T_i} = 1|\mathrm{X_i}) < 1$. According to the positivity assumption, there is a non-zero probability that each individual gets either treatment. The positivity assumption is violated practically when some individuals obtain treatment essentially often (or nearly never) when $\mathrm{P}(\mathrm{T_i} = 1|\mathrm{X_i}) \approx 0$ and 1. Besides, $\mathrm{P}(\mathrm{T_i} = 1|\mathrm{X_i})$ is called the PS, $\mathrm{e_i}$.

PS ESTIMATION ON ENSEMBLE LEARNING APPROACHES

PS is proposed as "conditional probability of assignment to a particular treatment given a vector of observed covariates".[1] In many fields, there is growing attention to using the PS to compare treated and untreated groups relying on the set of variables and to decrease bias in estimating ATEs and average treatment effects on the treateds in observation studies.

Logistic Regression is widely utilized as a parametric method to estimate PS value and its equation is written by

$$\hat{\mathrm{e}}(\mathrm{x}) = \Pr(\mathrm{T} = \mathrm{t}|\mathrm{X} = \mathrm{x}, \gamma) = \frac{\exp(\mathrm{X^{tr}}\gamma)}{1+\exp(\mathrm{X^{tr}}\gamma)} = \gamma_0 + \gamma_1 \mathrm{X}_1 + \cdots + \gamma_n \mathrm{X}_n \quad (1)$$

Unfortunately, logistics regression needs to make assumptions about variable selection, its functional forms and distributions, and well-defined interaction and higher-order terms.

Bagging or Bootstrap Aggregation: It is proposed bootstrap aggregating, called bagging, which aims to lower the variance of a statistical learning process in order that increase the accuracy of predictions.[27] Bagging, one of the well-known tree-based learning techniques in machine learning. Moreover, the bagging method consists of two main steps: bootstrap and aggregating.

Random Forest (RF): Breiman proposes the random forest approach, which presents an automatic, non-parametric technique for dealing with regression issues such as complicated interaction terms, nonlinear relationships between covariates, or both of several variables on the outcome.[27]

Bootstrapping: At the beginning of the 1990s, one of the oldest references to ensemble methods is Schapire's boosting approach, which described how “strong” classifiers can be created by linearly combining several poor ones by iteratively re-weighting training inputs. Boosting algorithm is a type of stage-wise additive modeling, where the model is built so that each stage concentrates on fixing the errors of the preceding stage's model, with model weaknesses being measured by a loss function.[28,29] There are many different types of boosting algorithms in literature: AdaBoost, GBM, Extreme gradient boosting (XGBoost).

GBM: It is proposed a gradient boosting algorithm that constructs based on the stagewise additive models by iteratively fitting the main model with the gradient descent method. Even though RF basically takes the mean of all random ensemble’s learnings, the boosting technique incrementally accumulates new model estimates, which means that each new model training is dependent on the error of the whole ensemble formed up to that point.[30]

XGBoost: XGBoost is introduced as an alternative of GBM that is one of the highest performing used for supervised learning.[31] XGBoost, which offers parallel processing and boosting, is a scalable upgraded version of GBM.

MW METHOD

Li and Greene proposed the MW method as an alternative to 1:1 matching and IPTW to make more effective estimated treatment effects and reduce bias between treatment groups.[19] Because extreme PSs cause produce large weights when these weightings are examined under the IPTW and matching method cases. Let f(x) represents the marginal probability of variables X in the sample that includes both treatment and control groups. Then, f(x)h(x) can be used to express the density of the target population when h(x) is a predetermined tilting function of x describing the target population. If a specific treatment group is interested (i.e., it can be the treatment or control group when two group case is considered), the marginal density function for the t group symbolizes $f_t(x) = \Pr(X = x|T = t)$. Each treatment group is expressed as $f_1(x) = f(x)e(x)$ and $f_0(x) = f(x)e(x)$, respectively. The relevant weights $w_t(x)$ are described for each treatment group as:

$$w_{t=1}(x) \propto \frac{f(x)h(x)}{f(x)e(x)} = \frac{h(x)}{e(x)} \text{ , when } t = 1 \tag{2}$$

$$w_{t=0}(x) \propto \frac{f(x)h(x)}{f(x)(1-e(x))} = \frac{h(x)}{(1-e(x))} \text{ , when } t = 0 \tag{3}$$

According to its definition, the MW is $\frac{\min\{e(x),1-e(x)\}}{e(x)}$for a treated group and $\frac{\min\{e(x),1-e(x)\}}{1-e(x)}$ for an untreated group. Thus, MW is formulated for each subject as:

$$w_i = \frac{\min\{e(x),1-e(x)\}}{T_i e_i + (1-T_i)(1-e_i)} \tag{4}$$

The MW estimator is defined as

$$\hat{\Delta}= \frac{\sum_{i=1}^{n} w_i T_i Y_i}{\sum_{i=1}^{n} w_i T_i} - \frac{\sum_{i=1}^{n} w_i (1-T_i) Y_i}{\sum_{i=1}^{n} w_i (1-T_i)} \tag{5}$$

PS, matching weighting, and treatment effects are simultaneously estimated by resolving the below-estimating equations regarding to$\vartheta = \left(\mu_1, \mu_0, \gamma^{tr}\right)^{tr}$ as described in Lunceford and Davidian's methodology:

$$0 = \sum_{i=1}^{n} \emptyset_i(\vartheta) = \sum_{i=1}^{n} \begin{bmatrix} w(X_i, T_i, \gamma)T_i\left(Y_i - \mu_1\right) \\ w(X_i, T_i, \gamma)(1 - T_i)\left(Y_i - \mu_0\right) \\ S_\gamma(X_i, \gamma) \end{bmatrix} \qquad (6)$$

where w_i is defined relying on the γ as $w(X_i, T_i, \gamma)$. According to equation (6), first equation is corresponded to $\mu_1 = \frac{E(w_i T_i Y_i)}{E(w_i T_i)}$, while second equation is corresponded to $\mu_0 = \frac{E(w_i(1-T_i)Y_i)}{E(w_i(1-T_i))}$. Then, third equation is generated based on the PS model for γ (i.e., equation (1)). Thus, MW estimator might define as $\hat{\Delta}= \mu_1 - \mu_0$.This technique's estimator represents M-estimator with an asymptotically normal distribution because of having unbiased with corresponding to each equation of (6).[4]

SIMULATION DESIGN

Predictors generation: This paper used some part of simulations that investigated the efficacy of several scenarios for application with MW as the foundation for the structure of Monte Carlo simulations.[32,33] As in the previous study, the data sets were generated with ten variables (X_1-X_{10}), dichotomous treatment assignment (T) with pr(T)=0.5, and a dichotomous outcome with pr(Y)=0.02. However, this paper is considered to added seven more covariates (i.e., X_{11}-X_{17}), which express as distractor variables, on previous work. I generated the seventeen covariates X_1-X_{17} for each of N individual. Standard normal distributions are used as the continuous variables and dichotomized forms of standard normal distribution parameters are used as the binary variables. X_2, X_4, X_7, X_{10}, X_{12}, X_{15}, X_{17} covariates are drawn from a standard normal distribution, while X_1, X_3, X_5, X_6, X_8, X_9, X_{11}, X_{13}, X_{14}, X_{16} covariates represent as dichotomized variables. Secondly, some weak or strong correlations between covariates are constructed such as $\mathrm{corr}(X_1, X_5) = 0.2$,$\mathrm{corr}(X_2, X_6) = 0.9$,$\mathrm{corr}(X_3, X_8) = 0.2$,$\mathrm{corr}(X_4, X_9) = 0.9$, $\mathrm{corr}(X_7, X_{11}) = 0.2$, $\mathrm{corr}(X_6, X_{12}) = 0.9$, $\mathrm{corr}(X_9, X_{16}) = 0.9$,$\mathrm{corr}(X_{10}, X_{17}) = 0.2$. To sum up, X_1-X_4 are associated with both treatment and outcome assignments even though X_{11}-X_{17} are hold any association with neither treatment nor outcome assignment. In addition, some of distractor variables only are constructed weak or strong association with main covariates.

Treatments generation: Four scenarios of the PS model was employed to create treatment assignments. All four scenarios have the formula of $\Pr(T = 1|X) = \frac{1}{\{1+\exp(-\text{scenario\_tr}-\delta\varsigma)\}}$, where each scenario is a function of the confounding factors that determined how complex the relationship between all factors and treatment assignment are. The factor $\delta \sim N(0,1)$ and ς factor takes value of 1. There are four versions for creating treatment assignments as follows:

$scenario\_tr_A$=$\beta_0 + \beta_1X_1$+β_2X_2+β_3X_3+β_4X_4+β_5X_5+β_6X_6+β_7X_7 (main effects terms)

$scenario\_tr_B$=$\beta_0 + \beta_1X_1$+β_2X_2+β_3X_3+β_4X_4+β_5X_5+β_6X_6+β_7X_7+$\beta_1X_2^2$+$\beta_4X_4^2$+$\beta_7X_7^2$ (main effects terms plus three quadratic terms)

$scenario\_tr_C$=$\beta_0 + \beta_1X_1$+β_2X_2+β_3X_3+β_4X_4+β_5X_5+β_6X_6+β_7X_7+$\beta_1\ X_1X_3$+$\beta_2X_2X_4$+ $\beta_3\ X_3X_5$+$\beta_4X_4X_5$+ $\beta_5X_5X_6$+ $\beta_5\ X_5X_7$ + $\beta_1\ X_1X_6$+$\beta_2X_2X_3$+$\beta_3\ X_3X_4$+$\beta_4X_4X_6$ (main effects terms plus ten interaction terms)

$scenario\_tr_D$=$\beta_0 + \beta_1X_1$+β_2X_2+β_3X_3+β_4X_4+β_5X_5+β_6X_6+β_7X_7+$\beta_1\ X_1X_3$+$\beta_2X_2X_4$+$\beta_3\ X_3X_5$

+$\beta_4X_4X_5$+ $\beta_5X_5X_6$+ $\beta_5\ X_5X_7$ + $\beta_1\ X_1X_6$+$\beta_2X_2X_3$+$\beta_3\ X_3X_4$+$\beta_4X_4X_6$+$\beta_2X_2^2$+$\beta_4X_4^2$+$\beta_7X_7^2$ (main effects terms plus three quadratic terms plus ten interaction terms)

The parameters$\beta = \{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7\}$ $= \{0, 0.8, -0.25, 0.6, -0.4, -0.8, -0.5, 0.7\}$. Then, a uniform distribution is used to create a random value ranging from 0 and 1. The value of T is adjusted to be 1 if the estimated PS value P(T=1|X) is greater than the randomly produced uniform value. Otherwise, it is adjusted as 0 if randomly generated uniform values are larger than the estimated PS values. To study a scenario like a two-arm non-randomized controlled trial, I intended to expose around 50 percent of the individuals to the treatment. The value of $\beta_0 = 0$, which is taken place generating treatment scenarios, is decided to assign the treatment to roughly half of the subjects (i.e., 50 percent treatment). However, the value of $\beta_0 = 2$ is set to generate 20 percent of the treated subject while the value of $\beta_0 = -2$ is set to provide that treatment consisted of approximately 80 percent of treated individuals. Thus, I create low, middle, and strong treatment probability across four treatment scenarios (i.e., treatment A, treatment B, treatment C, and treatment D). Thus, I obtain low, middle, and strong treatment probability across four treatment scenarios (i.e., treatment A, treatment B, treatment C, and treatment D).

Outcome generation: A scenario of the PS model is used to obtain outcome assignments. The outcome scenario has the formula of $\Pr(\mathrm{Y} = 1|\mathrm{T}, \mathrm{X}) = \frac{\exp(\text{scenario\_out}+\delta\mathrm{T})}{\{1+\exp(\text{scenario\_out}+\delta\mathrm{T})\}}$, where version of the outcome is defined as a function of complex model and also, true treatment δ is represented by -0.4 meanwhile different treatment scenarios T is replaced in place.

$scenario\_out = \alpha_0 + \delta_1 T + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_8 + \alpha_6 X_9 + \alpha_7 X_{10}$(main effect terms)

The parameters $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$= $\{0.3, -0.36, -0.73, -0.2, 0.71, -0.19, 0.26, -0.4\}$.

Moreover, intercept term α_0 is set to be constant value with -3.85, while true treatment effect δ_1 set to be 0.4 value. Like generating treatment assignment, the value of outcome Y is assigned to be 1 if $\Pr(\mathrm{Y} = 1|\mathrm{T}, \mathrm{X})$ value is larger than the randomly produced uniform value. Otherwise, the outcome Y is set to be 0.

The propensity estimation strategies: The goal of the Monte Carlo simulation is to ascertain which PS models are best at balancing the seventeen variables between treatment and outcome assignments. So, five PS model versions are taken into consideration, with each having a different selection of covariates used in the model. In each of the treatment and outcome cases, the functional version of the PS generating model consists of the factors listed below.[32]

The main PS model (model 1): $\mathrm{X}_1 - \mathrm{X}_{10}$ covariates, which relate to either outcome or treatment assignments, are used.

The true PS model (model 2): $\mathrm{X}_1 - \mathrm{X}_7$ covariates, which have a direct relationship to treatment assignments, are used.

The confounder model (model 3): $\mathrm{X}_1 - \mathrm{X}_4$ and $\mathrm{X}_8 - \mathrm{X}_{10}$ covariates, which are associated with the outcome assignments, are used.

The true confounder model (model 4): $\mathrm{X}_1 - \mathrm{X}_4$ covariates, which relate with the treatment and outcome assignments, are used.

The full model (model 5): All covariates $(\mathrm{e.i.}, \mathrm{X}_1 - \mathrm{X}_{17})$ are used.

Evaluation metrics for the simulation: First evaluation metric of simulation is the average standard error, which is the mean of a given ATE's 1000 standard errors, i.e.,

$$\text{The mean } \widehat{\mathrm{SE}} = \frac{1}{1000} \sum_{i=1}^{1000} \widehat{\mathrm{SE}}_i$$

Second metric across the simulation is the RMSE that denoted taking square root of means square error for each estimator, i.e.,

$$\mathrm{RMSE} = \sqrt{\frac{1}{1000} \sum_{i=1}^{1000} \left(\widehat{\mathrm{ATE}}_i - \mathrm{ATE}\right)^2}$$

where estimated average treatment effect $\widehat{\mathrm{ATE}}$ and true treatment exposure ATE are defined. Last used metric is bias, which is defined as the difference between the mean estimated treatment effect and the true effect set at -0.4, i.e.,

$$\mathrm{Bias} = \frac{1}{1000}\sum_{i=1}^{1000}\left(\widehat{\mathrm{ATE}}_{\mathrm{i}} - \mathrm{ATE}\right)$$

This study generated 1000 datasets with 1000 individuals in each simulation.

RESULTS

Monte Carlo simulation is used to assess how the variables selection for different PS model scenarios (i.e., Model-1, Model-2, Model-3, Model-4 and Model-5) across treatment (i.e., Treatment A, Treatment B, Treatment C and Treatment D) against considering different probabilities of each treatment scenario, and outcome scenarios impact the consistency of the MW treatment and control individuals. The average of estimated SE($\overline{\widehat{\mathrm{SE}}}$), bias and RMSE provide as summaries of the findings.

TABLE 1: Performance of SE and RMSE of Logistic Regression, Bagging, Random Forest, GBM, XGBoost propensity score methods across all propensity score models (i.e., Model1-5) in all treatment versions (i.e., Treatment A-D), where are P(Treatment) ≅0.5.

		Treatment A		Treatment B		Treatment C		Treatment D	
		SE	*RMSE*	*SE*	*RMSE*	*SE*	*RMSE*	*SE*	*RMSE*
Logistic Regression	*Model 1*	2.275	0.563	1.413	0.553	0.582	0.547	0.512	0.165
	Model 2	2.226	0.564	1.384	0.551	0.548	0.547	0.482	0.164
	Model 3	1.054	0.548	0.897	0.550	0.561	0.547	0.494	0.165
	Model 4	1.033	0.547	0.887	0.548	0.521	0.547	0.457	0.164
	Model 5	2.415	0.564	1.490	0.555	0.630	0.548	0.557	0.165
Bagging	*Model 1*	0.981	0.565	1.224	0.559	0.656	0.546	0.570	0.168
	Model 2	0.951	0.565	1.198	0.563	0.500	0.546	0.439	0.164
	Model 3	0.840	0.553	0.952	0.554	0.649	0.548	0.564	0.167
	Model 4	0.798	0.550	0.914	0.553	0.495	0.546	0.434	0.164
	Model 5	1.059	0.560	1.283	0.562	0.758	0.551	0.663	0.173
Random Forest	*Model 1*	0.895	0.557	1.117	0.552	0.621	0.546	0.540	0.165
	Model 2	0.909	0.555	1.135	0.548	0.592	0.547	0.519	0.164
	Model 3	0.784	0.549	0.889	0.547	0.624	0.547	0.545	0.165
	Model 4	0.710	0.544	0.760	0.543	0.518	0.547	0.454	0.164
	Model 5	0.858	0.553	1.057	0.551	0.612	0.546	0.535	0.165
GBM	*Model 1*	1.106	0.561	1.489	0.568	0.529	0.549	0.455	0.166
	Model 2	1.111	0.566	1.542	0.564	0.518	0.548	0.443	0.164
	Model 3	0.777	0.545	0.901	0.550	0.529	0.549	0.454	0.167
	Model 4	0.773	0.549	0.886	0.548	0.511	0.549	0.436	0.164
	Model 5	1.081	0.566	1.445	0.565	0.530	0.548	0.458	0.165
XGBoost	*Model 1*	0.901	0.570	1.080	0.555	0.757	0.546	0.664	0.172
	Model 2	0.901	0.571	1.079	0.556	0.553	0.546	0.484	0.167
	Model 3	0.847	0.557	1.018	0.549	0.760	0.549	0.665	0.172
	Model 4	0.837	0.553	1.017	0.551	0.534	0.547	0.469	0.166
	Model 5	0.906	0.570	1.086	0.558	0.802	0.547	0.702	0.175

SE: Standard error; RMSE: Root mean squared error; GBM: Gradient Boosting Machine; XGBoost: Extreme gradient boosting

Table 1 shows the $\overline{\widehat{\text{SE}}}$ and the RMSEs for various simulation scenarios when Pr(T)≅0.5 is considered generating treatment scenarios. The $\overline{\widehat{\text{SE}}}$ over Model-1 in the logistic regression model is 2.275 for Treatment A, 1.413 for Treatment B, 0.582 for Treatment C, and 0.512 for Treatment D, while the $\overline{\widehat{\text{SE}}}$ over model 4 in the logistic regression model is 1.033 for Treatment A, 0.887 for Treatment B, 0.521 for Treatment C and 0.457 for Treatment D. Similarly, the $\overline{\widehat{\text{SE}}}$ from logistic regression models' method are nearly 2.5 times higher than the ones from the RF methods across Model-1, Model-2, and Model-5 under Treatment A scenario. However, the difference in the $\overline{\widehat{\text{SE}}}$s in linear treatment scenarios (i.e., Treatment A) between methods have been more than the ones in complex non-linear treatment scenarios (i.e., Treatment D). In other words, among of models across methods, there is nearly no significant difference in terms of the $\overline{\widehat{\text{SE}}}$s and the RMSEs under the Treatment D. Overall, the findings state that logistic regression runs a poor performance than using any ensemble learning methods with linear treatment scenarios (i.e., Treatment-A) when there is no violation of the positivity assumption. RF yields nearly lowest the $\overline{\widehat{\text{SE}}}$ and the RMSE in all scenarios but the SE for logistic regression is highest in Treatment A, while GBM hold the lowest $\overline{\widehat{\text{SE}}}$ values in Treatment D across all models. Thus, it concludes that it significantly is important whether the treatment scenario is generated based on the linear or not linear forms for logistics regression than being ensemble learning.

TABLE 2: Performance of SE and RMSE of Logistic Regression, Bagging, Random Forest, GBM, XGBoost propensity score methods across all propensity score models (i.e., Model1-5) in all treatment versions (i.e., Treatment A-D), where are P(Treatment) ≅0.8.

		Treatment A		**Treatment B**		**Treatment C**		**Treatment D**	
		SE	*RMSE*	*SE*	*RMSE*	*SE*	*RMSE*	*SE*	*RMSE*
Logistic Regression	*Model 1*	272.56	2.309	157.66	1.872	1.915	1.503	1.162	0.841
	Model 2	250.98	2.308	144.49	1.870	1.710	1.494	1.034	0.834
	Model 3	145.75	2.334	89.094	1.870	1.746	1.495	1.066	0.835
	Model 4	135.63	2.334	88.577	1.869	1.502	1.486	0.916	0.829
	Model 5	333.33	2.308	162.13	1.870	2.253	1.524	1.357	0.853
Bagging	*Model 1*	73.426	2.255	60.833	1.811	2.136	1.501	1.264	0.857
	Model 2	68.391	2.258	60.798	1.812	1.329	1.475	0.819	0.826
	Model 3	67.813	2.266	39.854	1.819	2.084	1.504	1.243	0.867
	Model 4	59.554	2.277	37.998	1.824	1.298	1.473	0.803	0.826
	Model 5	86.777	2.238	61.828	1.803	2.531	1.456	1.518	0.850
Random Forest	*Model 1*	92.499	2.342	78.252	1.857	1.932	1.507	1.169	0.844
	Model 2	96.213	2.341	78.591	1.852	2.242	1.512	1.340	0.844
	Model 3	75.796	2.341	50.527	1.860	1.973	1.508	1.183	0.844
	Model 4	63.224	2.319	39.255	1.824	1.525	1.484	0.932	0.827
	Model 5	91.015	2.340	69.786	1.855	1.816	1.503	1.107	0.842
GBM	*Model 1*	90.118	2.334	60.829	1.849	1.393	1.469	0.853	0.821
	Model 2	79.849	2.341	53.896	1.844	1.420	1.459	0.867	0.820
	Model 3	59.477	2.306	35.304	1.838	1.386	1.466	0.850	0.822
	Model 4	63.569	2.326	36.328	1.842	1.423	1.458	0.852	0.820
	Model 5	69.492	2.304	58.915	1.835	1.389	1.464	0.848	0.829
XGBoost	*Model 1*	51.239	2.281	38.338	1.813	1.732	1.504	1.057	0.841
	Model 2	51.860	2.281	38.242	1.816	1.507	1.479	0.926	0.829
	Model 3	48.689	2.284	34.187	1.856	1.729	1.505	1.059	0.841
	Model 4	48.436	2.284	34.173	1.853	1.448	1.475	0.888	0.826
	Model 5	50.453	2.275	38.072	1.812	1.729	1.518	1.060	0.856

SE: Standard error; RMSE: Root mean squared error; GBM: Gradient Boosting Machine; XGBoost: Extreme gradient boosting

The $\widehat{SE}$ and RMSE as the treatment modeled association between exposure and covariate gets less complex, ensemble learning techniques (bagging, RF, GBM, and XGBoost) tended to yield less values than using logistic regression across all PS models versus all treatment scenarios when the violation of positivity assumption is considered in Table 2. In addition, there is a huge decreasing trend in $\widehat{SE}$ and RMSEs from Treatment-A to Treatment-D in all methods across all models. When considering more complex treatment scenarios (i.e., Treatment-D) in Table 2, GBM performs less $\widehat{SE}$ than the rest of the three ensemble learning methods and logistic regression methods across all PS models. Looking at Table 3, the $\widehat{SE}$s for logistic regression were remarkably higher for all treatment scenarios and across all five model scenarios in Treatment A. Like Table 2, there are remarkably decreasing intention for $\widehat{SE}$ and RMSEs from Treatment A to Treatment D, while the performance of GBM is generally smaller $\widehat{SE}$ in scenario of Treatment D with a $\widehat{SE}$s of 0.658,0.668,0.658,0.677 and 0.659 across PS models (i.e., Model 1-5) compared to the rest of methods under existing of positivity violation in Table 3.

TABLE 3: Performance of SE and RMSE of Logistic Regression, Bagging, Random Forest, GBM, XGBoost propensity score methods across all propensity score models (i.e., Model1-5) in all treatment versions (i.e., Treatment A-D), where are P(Treatment) ≅0.2.

		Treatment A		Treatment B		Treatment C		Treatment D	
		SE	*RMSE*	*SE*	*RMSE*	*SE*	*RMSE*	*SE*	*RMSE*
Logistic Regression	*Model 1*	319.44	2.585	174.2	2.42	1.001	0.500	0.830	0.106
	Model 2	314.21	2.581	185.2	2.42	0.917	0.495	0.764	0.099
	Model 3	121.57	2.577	64.7	2.35	0.935	0.493	0.776	0.103
	Model 4	116.53	2.574	62.9	2.35	0.831	0.488	0.695	0.097
	Model 5	361.90	2.589	201.2	2.43	1.160	0.512	0.955	0.122
Bagging	*Model 1*	83.371	2.558	68.5	2.32	1.098	0.519	0.893	0.148
	Model 2	80.355	2.563	72.5	2.33	0.744	0.485	0.626	0.092
	Model 3	72.541	2.558	60.7	2.31	1.077	0.508	0.882	0.147
	Model 4	66.417	2.561	57.9	2.32	0.730	0.484	0.617	0.091
	Model 5	94.664	2.550	75.7	2.31	1.313	0.520	1.071	0.159
Random Forest	*Model 1*	93.686	2.585	76.5	2.35	1.010	0.501	0.839	0.121
	Model 2	107.01	2.600	80.6	2.35	1.142	0.508	0.949	0.117
	Model 3	75.688	2.588	71.2	2.36	1.035	0.497	0.862	0.123
	Model 4	67.205	2.592	52.7	2.34	0.848	0.484	0.713	0.096
	Model 5	86.825	2.586	77.7	2.36	0.979	0.503	0.818	0.127
GBM	*Model 1*	93.849	2.590	71.8	2.35	0.787	0.497	0.658	0.103
	Model 2	72.248	2.603	79.8	2.35	0.793	0.490	0.668	0.096
	Model 3	64.384	2.596	50.8	2.35	0.785	0.496	0.658	0.099
	Model 4	54.739	2.592	48.9	2.35	0.797	0.486	0.677	0.099
	Model 5	99.183	2.613	92.9	2.36	0.784	0.499	0.659	0.102
XGBoost	*Model 1*	63.386	2.593	55.0	2.37	0.987	0.512	0.837	0.136
	Model 2	59.796	2.579	51.1	2.35	0.861	0.496	0.736	0.104
	Model 3	63.808	2.595	56.4	2.33	0.986	0.513	0.838	0.135
	Model 4	64.916	2.594	56.4	2.33	0.828	0.487	0.702	0.100
	Model 5	63.028	2.594	55.9	2.37	0.996	0.520	0.849	0.152

SE: Standard error; RMSE: Root mean squared error; GBM: Gradient Boosting Machine; XGBoost: Extreme gradient boosting

In principle, the plots of the bias throughout all methods reveal that the findings were comparable under the various PS estimating strategies as seen in Figure 1.

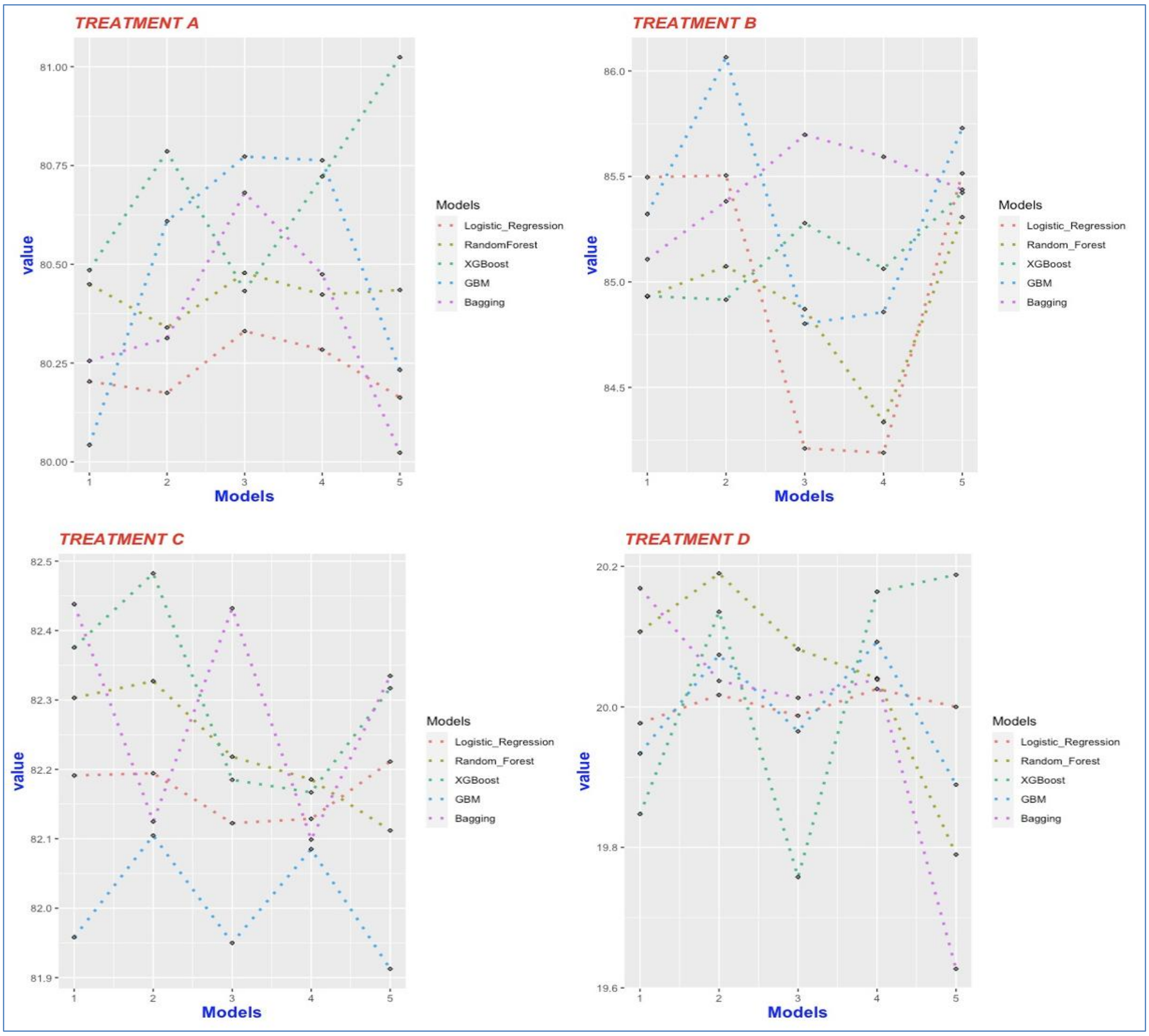

FIGURE 1: Performance of Bias of Logistic Regression, Bagging, Random Forest, GBM, XGBoost propensity score methods across all propensity score model scenarios (i.e., Model1-5) in all treatment versions (i.e., Treatment A-D), where are P(Treatment)≅0.5.

GBM: Gradient Boosting Machine; XGBoost: Extreme gradient boosting.

When treatment scenarios and PS model complexity were extremely low in terms of comparing bias (i.e., Treatment D and Model-5), bagging ad RF techniques outperforms the bootstrapping techniques and logistic regression method. Overall, less complex treatment scenarios (Treatment-A-B-C) are more likely to provide remarkably large biases estimates in all model scenarios across all methods.

DISCUSSION

Most statistical software might be used to conduct matching weighting, making it generally available to both statisticians and non-statisticians.[19,34] The purpose of this study is to emphasize thing: the scientist's requirement to adhere to an estimate strategic plan in which the estimand is properly specified before the estimation method is adopted. Even though researchers rapidly implement reliable PS estimates based on the logistics regression, and the super learner to use them within matching weighting method each with its advantages and limitations, there is no assessment implementation process of ensemble learning approaches to estimate PS before applying matching methods.[21-24,34] As result, when choosing a method, researchers may need to consider the complexity of the PS estimation model, the complexity of the generating treatment scenarios and under the violation of positivity assumption.[22,35] Along with these recommendations, I also offer a data-analytic methodology for selecting between the logistic regression and ensemble learning approaches.

CONCLUSION

The estimated PS based on the different ensemble learning method and logistics regression was used to determine whether the assumptions made when employing MW are sufficient for obtaining accurate causal inference. I outlined a through set of diagnostics to evaluate if weighting sample by matching weighting provided resulted in a sample where distribution of observed baseline variables was the same for treatment and control individuals.

- Ensemble learning techniques (bagging, RF, GBM, and XGBoost) tended to yield less the mean SE and RMSE than the logistic regression whatever the different probability generating of treatment across the four treatment models was considered.
- When the resulted are review from less complex treatment model (i.e., Treatment A) to complex treatment model, the resulted of SE and RMSE exhibit less values across same PS models under the same condition of the probability of treatment values.
- If positivity assumption was violated (i.e., in Table 2 and Table 3), all models across all treatment model scenarios illustrated that there was explosion the resulted of SE and RMSE for logistic regression under the linear treatment scenario (i.e., Treatment-A).

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

This study is entirely author's own work and no other author contribution.

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55. [Crossref]
2. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17(19):2265-81. [Crossref] [PubMed]
3. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci. 2010;25(1):1-21. [Crossref] [PubMed] [PMC]
4. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med. 2004;23(19):2937-60. Erratum in: Stat Med. 2017;36(14):2320. [Crossref] [PubMed]
5. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. Review of Economics and Statistics. 2004;86(1):4-29. [Crossref]
6. Tipton E. Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts. Journal of Educational and Behavioral Statistics. 2013;38(3):239-66. [Crossref]
7. Austin PC, Stuart EA. 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. 2015;34(28):3661-79. [Crossref] [PubMed] [PMC]

8. Ertefaie A, Stephens DA. Comparing approaches to causal inference for longitudinal data: inverse probability weighting versus propensity scores. Int J Biostat. 2010;6(2):Article 14. [Crossref] [PubMed]
9. Sugihara M. Survival analysis using inverse probability of treatment weighted methods based on the generalized propensity score. Pharm Stat. 2010;9(1):21-34. [Crossref] [PubMed]
10. Garrido MM. Covariate adjustment and propensity score. JAMA. 2016;315(14):1521-2. [Crossref] [PubMed] [PMC]
11. Zou B, Zou F, Shuster JJ, Tighe PJ, Koch GG, Zhou H. On variance estimate for covariate adjustment by propensity score analysis. Stat Med. 2016;35(20):3537-48. [Crossref] [PubMed] [PMC]
12. Imbens GW, Rubin DB. Causal Inference in Statistics, Social, and Biomedical Sciences. 1st ed. Cambridge: Cambridge University Press; 2015. [Crossref]
13. Keele L, Lenard M, Page L. Overlap violations in clustered observational studies of educational interventions. Journal of Research on Educational Effectiveness. 2022:1-18. [Crossref]
14. Zhu Y, Hubbard RA, Chubak J, Roy J, Mitra N. Core concepts in pharmacoepidemiology: violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. Pharmacoepidemiol Drug Saf. 2021;30(11):1471-85. [Crossref] [PubMed] [PMC]
15. Zhu Y, Mitra N, Roy J. Addressing positivity violations in causal effect estimation using gaussian process priors. arXiv. 2021. [Crossref] [PubMed]
16. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009;96(1):187-99. [Crossref]
17. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution--a simulation study. Am J Epidemiol. 2010;172(7):843-54. [Crossref] [PubMed] [PMC]
18. Walker AM, Patrick AR, Lauer MS, Hornbrook MC, Marin MG, Platt R, et al. A tool for assessing the feasibility of comparative effectiveness research. Comp Eff Res. 2013;3:11-20. [Crossref]
19. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. Int J Biostat. 2013;9(2):215-34. [Crossref] [PubMed]
20. Austin PC. Bootstrap vs asymptotic variance estimation when using propensity score weighting with continuous and binary outcomes. Stat Med. 2022;41(22):4426-43. [Crossref] [PubMed] [PMC]
21. Matsouaka RA, Zhou Y. A framework for causal inference in the presence of extreme inverse probability weights: the role of overlap weights. arXiv. 2020. [Link]
22. Zhou Y, Matsouaka RA, Thomas L. Propensity score weighting under limited overlap and model misspecification. Stat Methods Med Res. 2020;29(12):3721-56. [Crossref] [PubMed]
23. Yoshida K, Hernández-Díaz S, Solomon DH, Jackson JW, Gagne JJ, Glynn RJ, et al. Matching weights to simultaneously compare three treatment groups: comparison to three-way matching. Epidemiology. 2017;28(3):387-95. [Crossref] [PubMed] [PMC]
24. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. Journal of the American Statistical Association. 2018;113(521):390-400. [Crossref]
25. Rubin DB. Using multivariate matched sampling and regression adjustment to control bias in observational studies. Journal of the American Statistical Association. 1979;74(366a):318-28. [Crossref]
26. Rubin DB. Bias reduction using Mahalanobis-metric matching. Biometrics. 1980;36(2):293-8. [Crossref]
27. Breiman L. Bagging predictors. Machine Learning. 1996;24(2):123-40. [Crossref]
28. Schapire RE. A brief introduction to boosting. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. 1999:1401-6. [Link]
29. Freund Y, Schapire RE. Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference. 1996:148-56. [Link]
30. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 2001;29(5):1189-232. [Crossref]
31. Osman AIA, Ahmed AN, Chow MF, Huang YF, El-Shafie A. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal. 2021;12(2):1545-56. [Crossref]
32. Setodji CM, McCaffrey DF, Burgette LF, Almirall D, Griffin BA. The right tool for the job: choosing between covariate-balancing and generalized boosted model propensity scores. Epidemiology. 2017;28(6):802-11. [Crossref] [PubMed] [PMC]
33. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol Drug Saf. 2008;17(6):546-55. [Crossref] [PubMed] [PMC]
34. Zhou T, Tong G, Li F, Thomas LE, Li F. PSweight: an R package for propensity score weighting analysis. ArXiv. 2010. [Crossref]
35. Mao H, Li L, Greene T. Propensity score weighting analysis and treatment effect discovery. Stat Methods Med Res. 2019;28(8):2439-54. [Crossref] [PubMed]