Journal of Statistical Research *https://doi.org/10.3329/jsr.v55i2.58809*
2021, Vol. 55, No. 2, pp. 335-358 *ISSN 0256 - 422 X*

THE IMPACT OF VARIABLE OMISSION ON VARIABLE IMPORTANCE MEASURES OF CART, RANDOM FOREST, AND BOOSTING ALGORITHMS

W. HOLMES FINCH
Ball State University, Muncie, IN, USA
Email: whfinch@bsu.edu

SUMMARY

When researchers use statistical models to gain insights into various phenomena, they make a tacit assumption that most (if not all) of the important predictor variables with respect to the outcome are included in the analysis. However, in practice this may not always be possible, whether because some important variables could not be measured, or because a researcher was not aware of all such important predictors. Prior research has shown that when important variables are omitted from both linear and nonlinear regression models, the model coefficients can be biased, with greater levels of bias being associated with larger correlations between the missing and retained variables. However, very little work has examined how such omissions impact the performance of variable importance measures used with popular machine learning algorithms. Therefore, the purpose of this simulation study was to address this gap in the literature and thereby provide insights into the impact of such omissions on variable importance measures for classification and regression trees, random forests, and boosting algorithms. Results showed that when an important variable is omitted from an analysis, other predictors that are correlated with and/or involved in an interaction with it will have inflated variables importance measures themselves. An empirical example and implications of these results are discussed.

Keywords and phrases: Regression models; Machine learning algorithms; Missing values;

1 Introduction

Machine learning algorithms are increasingly important tools in a variety of research areas. Researchers in fields as diverse as education (Hew et al., 2018), business (Christensen et al., 2017; Zhu et al., 2019), medicine (Dias et al., 2019; Rossi et al., 2018), psychology (Delgadillo and Gonzalez Salas Duhne, 2020; Bone et al., 2016), and engineering (Zhang and Zhang, 2016; Gautier et al., 2015), to name a few, have made use of these tools in order to better understand processes underlying a variety of phenomena, and to obtain predictions for outcome variables of interest. These models come in a variety of forms, including linear and logistic regression, recursive partitioning algorithms, ensemble learning techniques, and neural networks. These models differ from one another in many 19 ways, including with the type of outcome data for which they are appropriate (e.g.,

dichotomous, counts, continuous scores), distributional assumptions made about the outcomes, and the nature of the relationships among the predictors. One commonality among these models is a general structure that includes an outcome variable of interest, and a set of independent or predictor variables that are thought to explain a relatively large portion of the variance in the outcome. In addition, these models also have in common a tacit assumption that the set of predictors used in the analysis contains all of the most important variables for predicting the outcome. In other words, there is an assumption that no important predictors have been left out of the model.

Although in an ideal case the predictive model does include all relevant independent variables for understanding the response, in practice it may not be realistic to make such an assumption. In some cases, researchers will know that such variables exist, but will not have access to them, while in other instances they may not be aware of some important factors that 4 should have been measured and included in the model. In either scenario, whether this tacit assumption is met falls to the data analyst and is dependent on exigencies within the domain of study. Prior research has found that the omission of important variables yielded biased coefficients in regression models, (Nystrom et al., 2019; Kutner et al., 2005). This bias in regression estimates was greater when the correlations among the predictor variables were higher. In addition, the Type I error rate for regression coefficients were inflated when the predictor correlations were higher.

The current Monte Carlo simulation study was designed to extend this earlier research by investigating how omission of important variables impacts importance measures associated with three widely used machine learning algorithms, classification and regression trees (CART), random forest (RF), and boosting. The remainder of the manuscript is organized as follows. First, a brief review of each machine learning method is presented, followed by a discussion of the literature on variable omission in the context of regression. The goals of the current study, and associated research hypotheses are then presented, followed by a description of the methods used to address them. Finally, the results of the simulation study are described, along with an empirical example, and then the results are discussed in the context of prior research with implications for practice.

1.1 Classification and regression trees

Classification and regression trees (CART) are based on a binary recursive partitioning algorithm designed to identify splits within a set of predictor variables that will optimally allocate observations in a sample into maximally separated groups based on values of a response variable (Breiman et al., 1984). The algorithm creates a series of binary splits (or nodes) based upon the predictors, thereby yielding a decision tree that culminates in a 5 set of terminal nodes for which the individuals contained therein cannot be further divided using the predictors. CART can be used with response variables of varying types, including continuous scores, ordinal, nominal, or dichotomous categorical variables. In order to achieve maximal separation based upon the response variable, the recursive partitioning algorithm finds the split among the predictor variables that minimizes the deviance within each node with respect to the outcome variable. For a continuous outcome this is akin to finding the split from among all of the predictors at a given node that minimizes the variance of the outcome within each of the resulting two nodes. A particular advantage for researchers using CART is that it doesn't rest on any assumptions regarding relationships between the predictors and

the outcome, and thus easily accommodates (and identifies) interactions among the predictors with respect to the response variable.

When interpreting the results of a CART analysis, researchers will not only want to understand the nature of the splits, but also the relative importance of each predictor variable in terms of building the tree. The measure of relative importance in this context was first introduced in Breiman et al. (1984), and then further described in Zhang and Singer (2010) . For each of the J variables at each candidate split, each of the predictor variables is assessed in turn. The optimal split for variable j is determined, and the resulting improvement in prediction of the outcome variable is then recorded. This is repeated for each of the J variables, and the optimal variable split is determined, and the algorithm continues on. However, all of the candidate splits and prediction accuracy values are recorded and then averaged across the tree for all of the variables. These individual importance measures are then summed, and the average importance across all splits is calculated for each variable and then divided by the sum. This proportion serves as the measure of relative importance for each variable.

1.2 Random forests

Random forest (RF) is an ensemble methodology, which applies CART to a dataset a large number (e.g., 500) of times, and then averages across the resulting trees to gain an understanding of how the response variable and predictors are related (Friedman et al., 2001). RF creates its tree ensemble by drawing B random samples of a subset (e.g., 75%) from the original sample, as well as a randomized subset of m predictors. CART is then applied to the subset of individuals and variables in order to grow a tree. For a given tree, the portion of the sample not used by the algorithm is known as the out of bag (OOB) sample, and can be used to validate the model, or for determining variable importance, as described below. The random sampling of both individuals and variables for each tree helps to minimize the overfitting problem endemic to CART, and also ensures that no single predictor variable dominates the overall model (Dietterich, 2000). RF has been found to successfully counter the tendency of CART to overfit the training data, thus improving the generalizability of predictions obtained from the (Breiman, 2001).

Variable importance in the context of RF is based upon a measure of improvement in the homogeneity of nodes created at a split. Thus, for each potential split of the Bth trees, each of the predictor variables is assessed using the bootstrap sample. The resulting split is then applied to the OOB sample and prediction accuracy for these cases is recorded. The effectiveness of the split is measured as the improvement in prediction accuracy from using the variable at that split. The values of the variable are then randomly permuted and the split is applied to the data. The difference in prediction accuracy as a result of randomly reorganizing the data, as compared to the accuracy for the original data, serves as the measure of variable importance. A larger difference between the results for the original and permuted data indicates that the variable is 7 relatively more important at that split, whereas a small difference would indicate that the variable is relatively less important. These values are then averaged across all m trees in the forest. As with CART, the variable importance values are then summed across the variables, and importance for a specific variable for the full forest is expressed as the ratio of its importance to the sum.

1.3 Boosting

Like RF, boosting is an ensemble method that combines a set of prediction models, such as trees, in order to develop a better predictor than would be possible using a single one. It is based on the principle of combining a set of weak learners in order to obtain a strong learning algorithm that can accurately predict the outcome of interest. In the first step of this process, an algorithm, such as CART (but any prediction model could be used) is applied to the data and predictions of the response variable are obtained for each individual in the sample. The difference between the predicted and actual values, known as the residuals, are then calculated. The model is then fit to these residuals and new prediction are made, and residuals are again calculated (Freund and Schapire, 1997; James et al., 2013). This process is repeated many times (e.g., 100), resulting in a model that should yield accurate predictions of the outcome through the gradual reduction of error, in the form of the residuals. The boosting approach has been shown to yield more accurate predictions of the outcome variable than do individual models, such as a single CART (Bauer and Kohavi, 1999).

Variable importance in the context of boosting is very similar to the approaches used with CART and RF. For each tree, the improvement in prediction accuracy is recorded for each variable at each split, and then these values are squared and averaged across all of the trees produced by the boosting algorithm. The sum across all such importance measures is then taken, 8 and the ratio of each individual variable and the sum is calculated in order to obtain a relative importance value. An alternate version of this method for assessing variable importance is to include the permutation of observations described above with respect to RF. In this version, all of the observations are permuted, and the split for variable j is applied. The difference in prediction accuracy for the split using the original data and that using the permuted data serves as the variable importance measure at that split. As with the other measure of importance, values are averaged across the splits for each variable and then divided by the sum of all variable importance measures across the trees produced by the boosting algorithm.

1.4 Prior research on omitted variables

There has been relatively little work examining the impact of omitting important variables from machine learning algorithms such as CART, RF, or boosting. However, there has been research investigating the impact of such omissions on the performance of linear and nonlinear regression models. Prior research in the context of linear regression models has demonstrated that when variables that are associated with the dependent variable in the population are omitted from the sample model, the result can be biased regression coefficients, poorly estimated standard errors, and an inflation of the Type I error rate (Afshartous and Preston, 2011; Kutner et al., 2005; Hölters et al., 2002). This work was extended to generalized linear models, and similar results with respect to parameter estimation bias and Type I error inflation were found (Cramer, 2005). Collectively, results of these studies showed that the degree of estimation bias was more severe when the omitted variable was strongly associated with the dependent variable. Furthermore, when the omitted variable was strongly correlated with one or more of the other predictors, parameter estimation bias was greater, and Type I error rates were more severely inflated. furthered this area of investigation by examining the impact

of omitting important independent variables when interactions are present in the data generating model. Their study involved a regression model in which the outcome variable was generated from the normal distribution, and was related to a set of two predictors through a nonlinear equation. The nonlinearity took the form of an interaction between two predictors. In addition to manipulating the magnitude of this interaction, Nystrom et al. (2019) . also examined the impact of the correlation between the independent variables. The goal of this study was to assess the impact of omitting one of these variables on the estimation of the regression relationship for the non-omitted variable. The results of this simulation study showed that the coefficient for the non-omitted variable exhibited positive bias that increased concomitantly with increases in the correlation with the omitted variable. In addition, this bias in the coefficient for the non-omitted variable decreased as the coefficient associated with the interaction term became a larger negative number. In summary Nystrom et al. (2019) found that there was larger bias present in the coefficient of the non-omitted variable when its relationship with the omitted variable was stronger, regardless of the population relationship between the omitted variable and the outcome. This bias was diminished as the coefficient for the interaction term was made more strongly negative.

1.5 Study goals

The purpose of this study was to investigate the performance of variable importance measures for popular and effective machine learning algorithms when an important variable in the population was omitted from the sample based data analysis. As reviewed above, prior work with linear and nonlinear regression models has demonstrated that such an omission can result in biased parameter estimates, incorrect standard errors, and inflated Type I error rates (Nystrom et al., 2019).Thus, it was of some interest to ascertain what impact such omissions might have on the variable importance assessment for other prediction models that are frequently used by researchers. The Monte Carlo simulation design, which is described below, focused on a relatively complex model that is based on one seen with actual data. It is hypothesized that omission of an important variable will alter the variable importance measures of a variable associated with it, and that this impact will be more severe when the correlation between the omitted and included variables is larger. A second hypothesis is that when the omitted variable is involved in an interaction with an included variable, the relative importance of the included variable will be impacted when a variable is omitted. A third hypothesis is that the relative importance of variables not associated with the omitted variable, either through correlation or interaction will be less severely impacted than is that of the variable associated with the omitted predictor.

2 Methods

In order to address the study goals described above, a Monte Carlo simulation methodology was used. The study conditions, which are described below, were drawn from prior research involving omitted variables in regression analysis, as well as empirical data to which machine learning algorithms have been applied. In addition to the simulation methodology, an empirical example was also

included in the study, and which is described in the Results section, below. For each combination of the study conditions described below, 1000 replications were done.

The data were generated using the following population model

$$y = \beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5+\beta_{12} x_1 x_2+\beta_{34} x_3 x_4+\beta_{123} x_1 x_2 x_3+\beta_{52} x_5^2+\epsilon. \quad (2.1)$$

Table 1: Data generating model parameter values

Model term	Value
β_0	0
β_1	1
β_2	0.5
β_3	0.5, 1, 1.5
β_4	0.2
β_5	0.7
β_{12}	0.4
β_{34}	0.2, 0.4, 0.8
β_{123}	0.2, 0.4, 0.8
β_{52}	0.4
ρ_{12}	0.4
ρ_{34}	0.2, 0.4, 0.8
ρ_{13}	0.2, 0.4, 0.8

The values of the model parameters appear in Table 1. The coefficients involving the target omitted variable, x_3 were manipulated in the study, as can be seen in the table. These values were selected to represent a range of effects from small to large for the main effects and interactions. The error term in model 3 was generated from a standard normal distribution (mean of 0, standard deviation of 1), as were each of the predictors. Given that the predictors and the error term were distributed as standard normal variates, the response variable was distributed as a standard normal as well. The R function `rnorm` was used to generate the predictors and error term. Equation 2.1 was then applied to these predictors in order to obtain a value of y for each observation in the simulated data. Finally, in addition to selected coefficients, the sample size was also manipulated to be 100, 500, or 1000.

Three models were fit to the data, CART, RF, and boosting. The `R` software package version 3.6.2 (R Core Development Team, 2018) was used. For CART, the `rpart` library was used, `party` was used to fit `RF`, and `gbm` was used for boosting. With respect to RF, 1000 trees were fit, with 75% of the sample and 3 of the predictors selected for each. For the boosted tree, a total of 1000 trees

for each analysis were fit to the data, and both the standard and permutation methods for calculating variable importance were used.

For each combination of sample size and parameter values, two models for each of the algorithms were fit to the data. The first of these models included all of the variables that appear in equation 2.1, and the second model excluded variable x_3 The outcome variables were the variable importance measures of x_4 for each of the algorithms. In order to ascertain which of the manipulated variables were related to the outcome, analysis of variance (ANOVA) was used in conjunction with a partial η^2 effect size. The ANOVA models included the manipulated factors and their interactions. Those that are found to be statistically significant were then examined in order to better understand the nature of these terms.

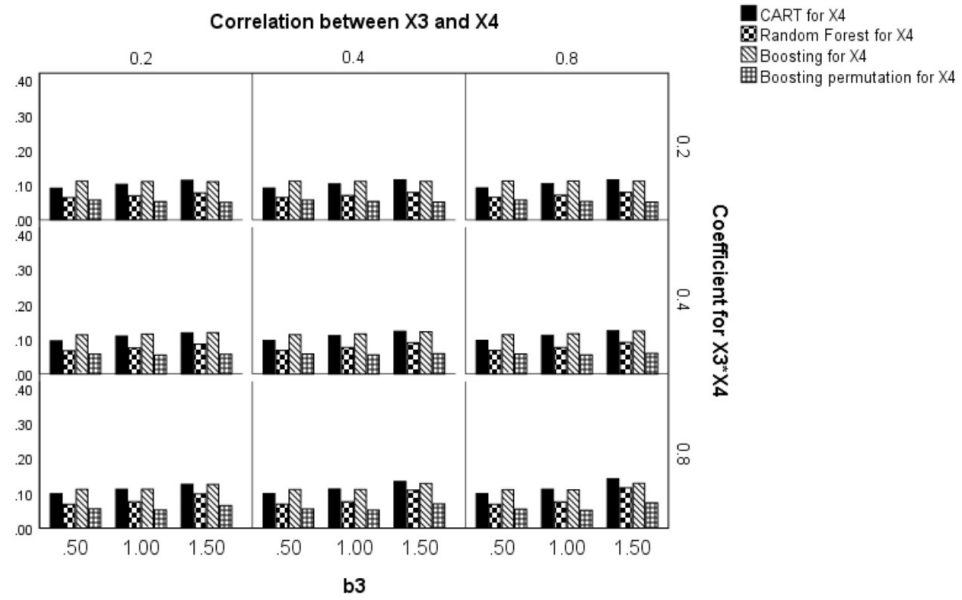
3 Results

3.1 Correct model

When the correct model was fit to the data, the highest order term that was statistically significantly related to the relative importance of X_4 was the interaction of method by the correlation between X_3 and X_4, by the coefficient of X_3*X_4 interaction by the coefficient of X_3 ($F_{24,336} = 6.035, p < 0.001, \eta^2 = 0.301$). All other terms were either not statistically significant, or subsumed in this interaction. Figure 1 Panel A displays the relative importance of X_4 method, correlation between X_3 and X_4, coefficient for X_3*X_4, and the coefficient of X_3, These results demonstrate that across prediction methods the relative importance for X_4 as generally similar across levels of the correlation with X_3 and across values of X_3. However, the importance values were slightly larger for larger coefficients of the X_3*X_4 interaction. With respect to the methods themselves, CART and the standard boosting algorithm consistently yielded the largest relative importance values, and the boosting approach based on permutation yielded the lowest relative importance for X_4.

Panel B of Figure 1 displays the relative importance of X_5 by prediction method, correlation between X_3 and X_4, X_3 and coefficient for the X_3*X_4 interaction. The purpose for including this variable was to assess how a variable completely unrelated to X_3 would be influenced by its inclusion and exclusion from the model. The results in Figure 1 Panel B reveal that the relative importance of X_5 as slightly lower for larger values of X_3, and that the two boosting approaches consistently yielded the largest relative importance values across the prediction methods. It was anticipated that as the relationship of X_3 increased in value, the relative importance ofX_5 would decline, given that its own relationship with the dependent variable remained unchanged. When compared with X_4 the relative importance of X_5 as greater, which might be expected given that X_5 and a population coefficient of 0.7 whereas X_4 had a coefficient of 0.2. Thus, even when X_4 as involved in a relatively large interaction effect with X_3 it was not found to be as relatively important as was X_5. One final point in this regard is that although X_5 had higher relative importance than X_4 for all prediction methods, this result was most notable for the boosting methods, whereas for CART and RF, the difference in relative importance was somewhat smaller.

Panel A: X4



Panel B: X5

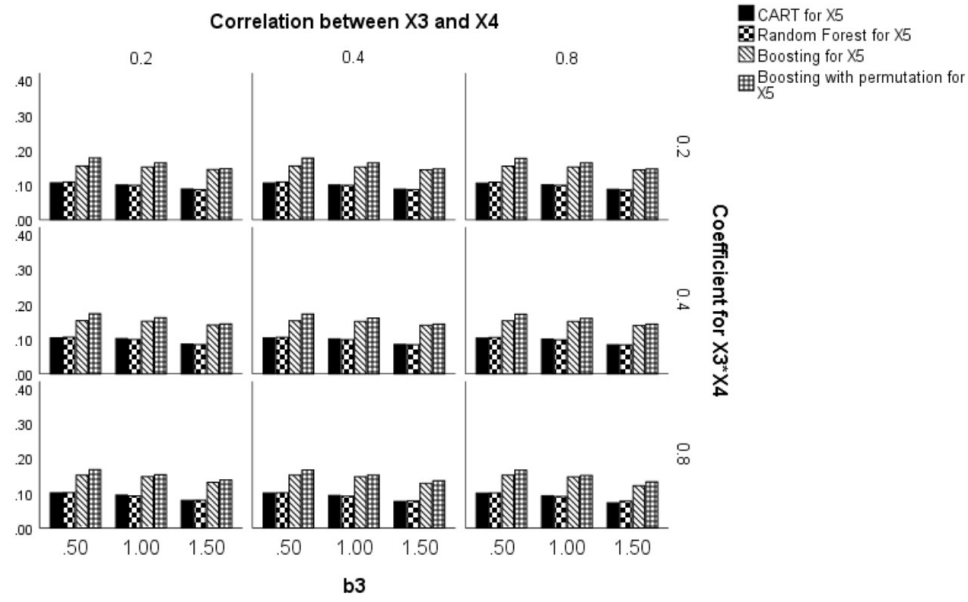


Figure 1: Relative importance of X_4 and X_5 by prediction method, correlation between X_3 and X_4, coefficient for $X_3 \times X_4$, and coefficient of X_3 : Correct model *(continued on next page)*

Panel C: X1

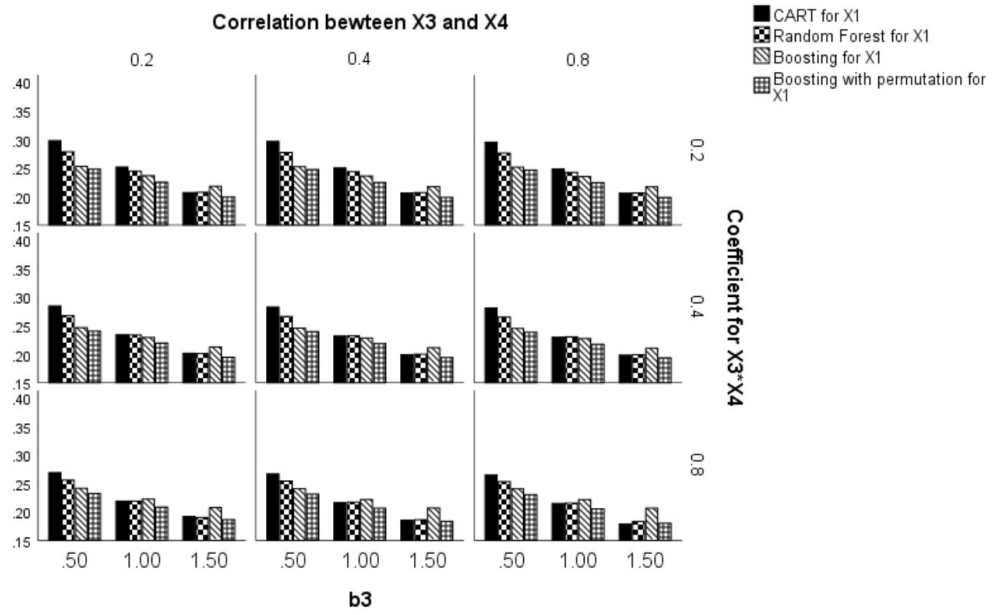


Figure 2: Relative importance of X_4 and X_5 by prediction method, correlation between X_3 and X_4, coefficient for $X_3 * X_4$, and coefficient of X_3 : Correct model

Finally, the relative importance values for variable X_1 appear in Figure 1 Panel C. Recall that X_1 was involved in a 3-way interaction with X_2 and X_3 and had a coefficient of 1 in the population. Across the other conditions, the relative importance of X_1 declined with increases in the value of the X_3 coefficient, and separately the coefficient for the X_3*X_4 interaction. There was relatively little relationship between the importance scores for X_1 and the correlation between X_3 and X_4.

3.2 Incorrect model

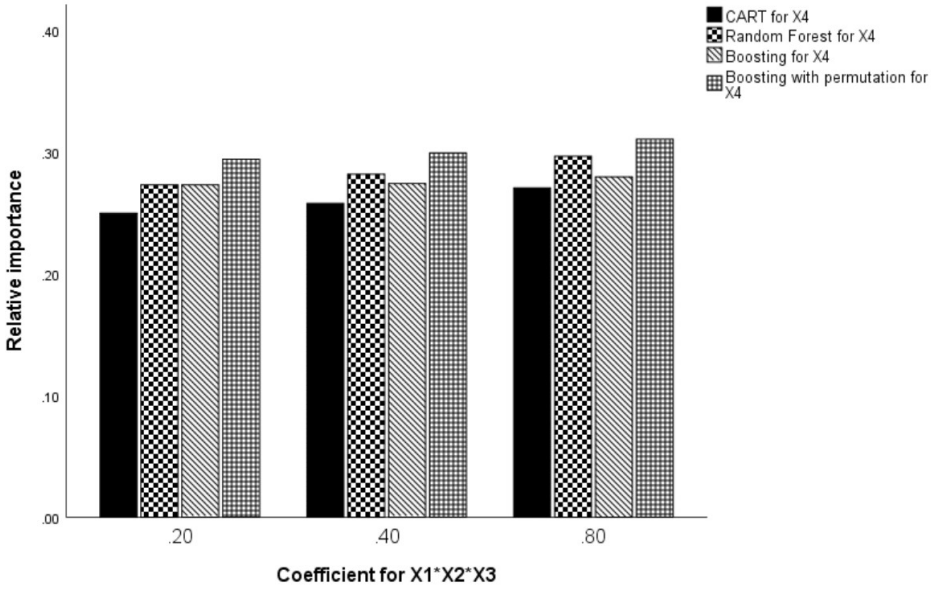
As was the case for the correct model including all of the independent variables that make up the prediction model, ANOVA was used to identify those factors that were associated with the relative variable importance of X_4 when X_3 was mistakenly removed from the model. The ANOVA results revealed that the interaction of prediction method by coefficient for the 3-way interaction of X_1*X_2*X_3 ($F_{6,222} = 34.137, p < 0.001, \eta^2 = 0.480$), and the interaction of method by the coefficient for X_3 by the coefficient for X_3*X_4 by the correlation between X_3 and X_4 ($F_{24,336} = 6.802, p < .001, \eta^2 = 0.327$) were both statistically significantly associated with the relative importance of X_4 All other manipulated variables in the simulation were either not associated with the relative importance of X_4, or were subsumed in one of these two interactions.

Figure 2 includes the relative importance forX_4 (Panel A), X_5 (Panel B), and X_1(Panel C) by prediction method and coefficient for the X_1*X_2*X_3 interaction. Perhaps the most notable result for both variables is that their relative importance was greater when X_3 was removed from the model than when it was included (see Figure 1). This would be expected, given that relative importance reflects the proportion of total importance accounted for by each variable in the model. Therefore, when there are fewer variables in the model, it would stand to reason that those remaining would each account for a larger share of the total importance; i.e., there are fewer variables among which the total importance must be shared. In addition, when X_3 is mistakenly omitted from the model, the relative importance of X_4 exceeds that of X_5 across conditions. Thus, X_4 went from having a relative importance value of between 0.07 and 0.15 when X_3 was in the model to between 0.25 and 0.35 when it was omitted. By contrast, the relative importance values for X_5 increased from between 0.10 and 0.20 when X_3 was in the model to between 0.18 and 0.25 when it was excluded. With respect to the interaction between method and the coefficient for X_1*X_2*X_3 the relative importance of $X_4$for each method increased concomitantly with increases in the coefficient value, with somewhat greater increases for CART and RF as compared to boosting. For X_1 or X_5 there was essentially no difference in the relative importance for each method across the interaction coefficient value.

Figure 3 contains the relative importance results for X_4 (Panel A), X_5 (Panel B), and X_1 (Panel C) by prediction method, the correlation between X_3 and X_4, the coefficient for the X_3*X_4 interaction, and the coefficient for X_3. As was evident in Figure 2, the relative importance of X_4 exceeded that of X_5 across study conditions, which was not the case when X_3 was correctly included in the model, and despite the fact that in the population X_5 had a larger coefficient than did X_4. In addition, the relative importance of X_4 increased concomitantly with increases in the population coefficient for X_3, and this effect was magnified by a larger correlation between X_3 and X_4. The relative importance of X_4 also increased in value along with increases in the correlation between X_3 and X_4, with the largest such effect occurring when the coefficient for the X_3*X_4 interaction was 0.8. Indeed, generally speaking the relative importance of X_4 was somewhat larger when the coefficient for its interaction with X_3 increased in value, which would be expected. With respect to the prediction methods, boosting coupled with the permutation approach yielded the largest importance values for X_4 across conditions, with RF providing similar results when the b3 coefficient was 1.5.

The relative importance of X_5 was largely unaffected by any of the study conditions, except for prediction method, where the permutation method used with the boosting algorithm yielded larger results than did the other methods. In contrast, the relative importance values for X_1 were lower when the value of X_3 was larger. In addition, as with both X_4 and X_5, the relative importance of X_1 was larger when X_3 was removed from the model. The difference in importance for X_1 between the two conditions was greater than was the case for X_5, but not as large as for X_3. This latter results seems to reflect the fact that whereas X_1 was part of a 3-way interaction with X_3, it was not correlated with X_3. In contrast, X_4 was involved in an interaction with X_3 and had a higher correlation value than did X_1 (which was 0).

Panel A: X4



Panel B: X5

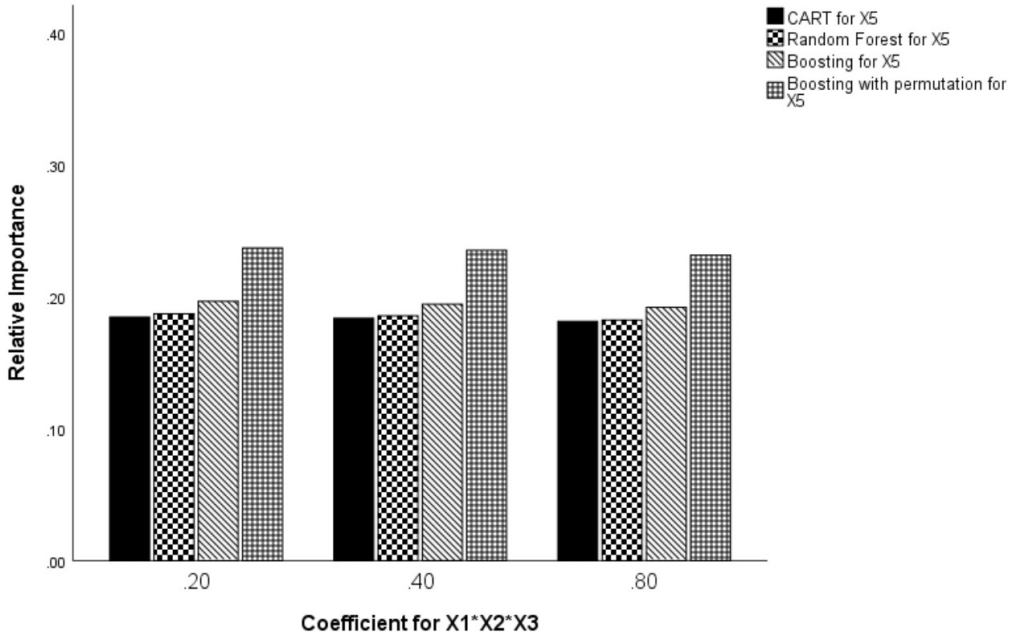


Figure 3: Relative importance of X_4 and X_5 by prediction method and coefficient of $X_1 * X_2 * X_3$ interaction: Incorrect model *(continued on next page)*

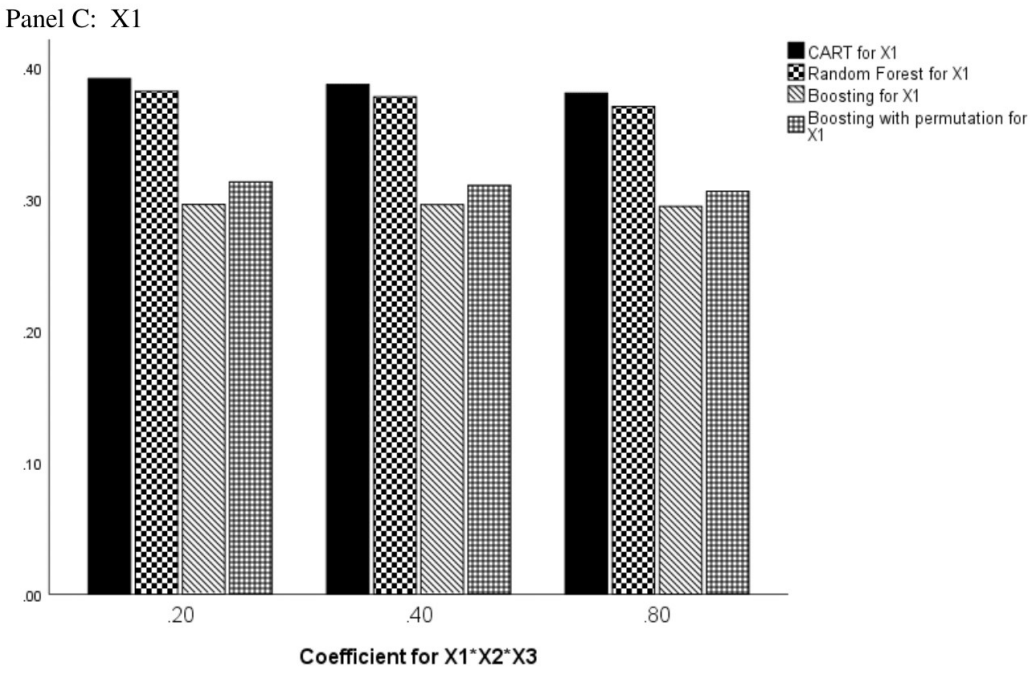
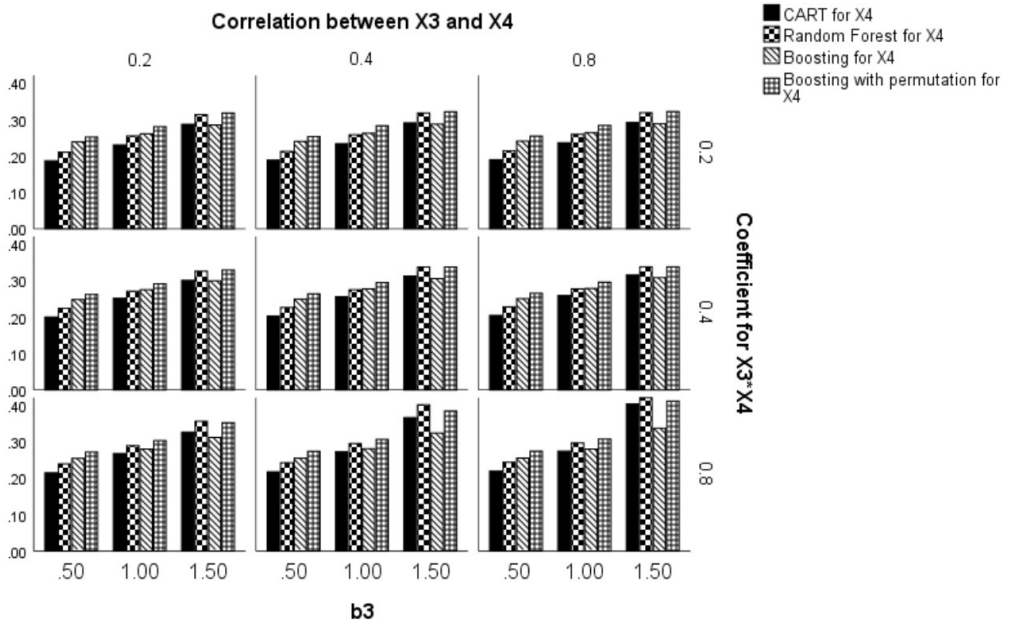


Figure 4: Relative importance of X_4 and X_5 by prediction method and coefficient of $X_1 * X_2 * X_3$ interaction: Incorrect model

Panel A: X4



Panel B: X5

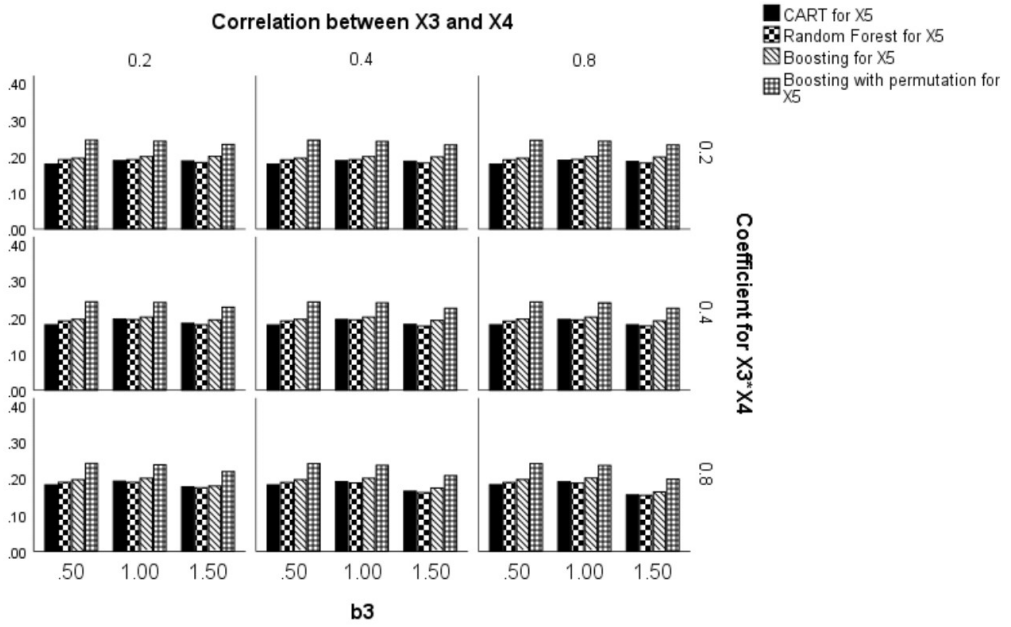


Figure 5: Relative importance of X_4 and X_5 by prediction method, correlation between X_3 and X_4, coefficient for $X_3 * X_4$, and coefficient of X_3 : Incorrect model *(continued on next page)*

Panel C: X1

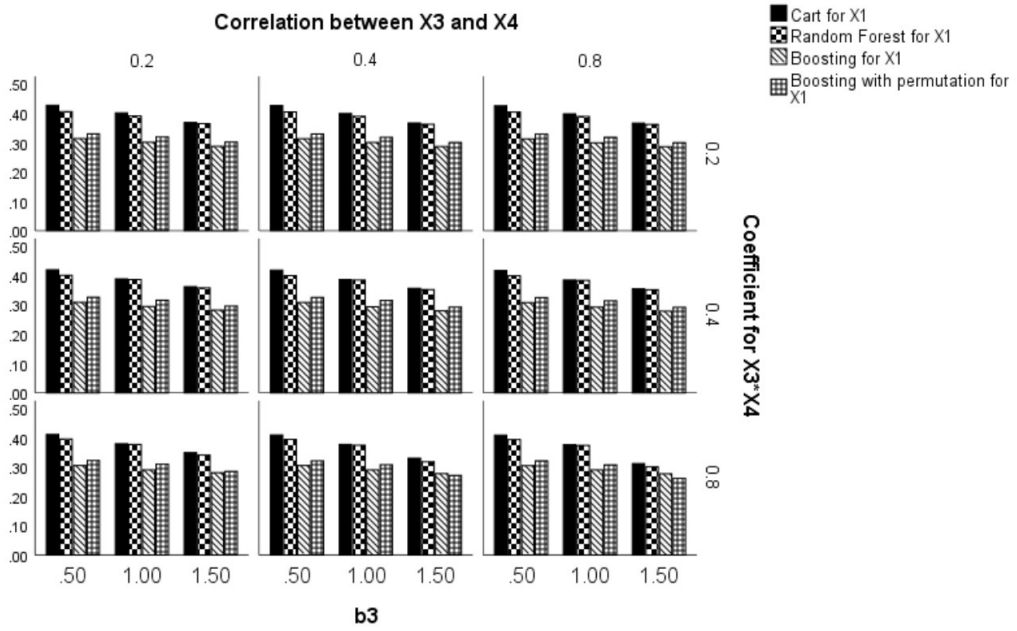


Figure 6: Relative importance of X_4 and X_5 by prediction method, correlation between X_3 and X_4, coefficient for $X_3 * X_4$, and coefficient of X_3 : Incorrect model

4 Empirical Example

In order to demonstrate the use of the relative importance measures, and the impact on them when important variables are removed from the model, data taken from the 2017 CDC Youth Risk Behavior Surveillance System (YRBS) were used (Centers for Disease Control and Prevention, 2018). The YRBS is a survey administered to 10th grade students in the United States that includes questions about a wide variety of health related behaviors. It includes items asking about student engagement in suicidal ideation and planning, violence, alcohol, tobacco, and drug use, dietary behaviors and physical activity, and sexual-related behaviors. The sample used in this study included 203, 663 participants from the 2017 administration of the survey. The outcome variable in this study was whether an individual had attempted suicide or not, with the predictor variables appearing in Table 2.

Table 2 includes the relative importance values for each prediction method when the full set of predictor variables was included. CART identified Considering suicide and Suffered physical dating violence as the two most important predictors of an adolescent making a suicide attempt. RF also identified the consideration of suicide as the most important predictor of a suicide attempt, along with being bullied at school. And as was true with CART, being the victim of physical dating

Table 2: Relative importance for suicide prediction model with all variables included

Variable	CART	RF	Boosting	Boosting permutation
Considered suicide	0.24	0.21	0.15	0.17
Physical dating violence	0.23	0.19	0.14	0.14
Sexual dating violence	0.15	0.07	0.04	0.03
Bullied at school	0.11	0.21	0.37	0.33
Bullied electronically	0.05	0.10	0.15	0.16
Current sexual activity	0.04	0.07	0.05	0.05
Carried gun last 12 months	0.03	0.04	0.02	0.03
Sexual identity	0.03	0.01	0.01	0.01
Weight loss recently	0.02	0.001	0.001	0.003
Sad/hopeless	0.02	0.02	0.01	0.02
Gender	0.02	0.004	0.002	0.003
Age	0.01	0.01	0.03	0.02
Initial alcohol use	0.01	0.01	0.002	0.004
Number of concussions	0.01	0.004	0.004	0.005
Amount of computer use per day	0.01	0.001	0.001	0.001
Fruit/vegetable consumption per day	0.004	0.01	0.01	0.01
Participate in sports team	0.004	0.03	0.01	0.01
Amount of sleep per night	0.002	0.003	0.004	0.004

violence was also an important predictor of a suicide attempt. Both relative importance measures associated with boosting identified being bullied at school as the most important predictor for attempting suicide. Considering suicide, being bullied electronically, and being the victim of physical dating violence followed being bullied at school in terms of relative importance.

In order to assess the relative importance of these variables when an important predictor was removed, each algorithm was fit to the data with the variable Considered suicide removed. The resulting relative importance measures for the remaining variables appear in Table 3. The relative importance results for CART appear to have changed the most when Considered suicide was removed from the analysis. In that case, the most important variable was Bullied at school, with a value of 0.51, compared to 0.11 and 4^{th} most important in the original analysis. The second most important variable was Sad/hopeless, which increased in relative importance from 0.02 to 0.16, and second most important as a predictor of attempting suicide. The physical dating violence variable fell in relative importance with a value of 0.13 (compared to 0.23 when Considered suicide was included). Finally, fewer variables had non-zero relative importance values when Considered suicide

Table 3: Relative importance for suicide prediction model with Considered Suicide removed

Variable	CART	RF	Boosting	Boosting permutation
Physical dating violence	0.13	0.18	0.17	0.15
Sexual dating violence	0.04	0.14	0.03	0.05
Bullied at school	0.51	0.33	0.40	0.38
Bullied electronically	0.06	0.07	0.18	0.20
Current sexual activity	0	0.07	0.04	0.05
Carried gun last 12 months	0	0.01	0.02	0.04
Sexual identity	0	0.001	0.02	0.01
Weight loss recently	0	0.01	0.003	0.003
Sad/hopeless	0.16	0.05	0.05	0.03
Gender	0	0.002	0.003	0.01
Age	0.01	0.02	0.04	0.03
Initial alcohol use	0	0.02	0.02	0.01
Number of concussions	0	0.004	0.01	0.01
Amount of computer use per day	0	0.001	0.001	0.01
Fruit/vegetable consumption per day	0.08	0.02	0.01	0.004
Participate in sports team	0	0.04	0.004	0.004
Amount of sleep per night	0	0.03	0.004	0.01

was removed from the analysis. The relative importance measures for RF, as well as both boosting based approaches appear to have been impacted less markedly than was true for CART. For these three methods, being bullied at school remained among the most, if not the single most, important predictor of a suicide attempt. In addition, being the victim of physical dating violence, and being bullied electronically also continued to have relatively high relative importance measures. There were some differences in relative importance among the less salient predictors for RF and boosting between the full and reduced models, but these differences were relatively minor. In summary, when Considered suicide was removed from the model the results for CART changed dramatically, with being bullied at school becoming far and away the most important predictor. In contrast, though individual importance values for the variables differed for RF and boosting with and without Considered suicide, these differences were not so great, and the relative ordering of the variables in terms of importance remained largely the same.

4.1 Identifying when important variables are omitted

The focus of this manuscript up to now has been centered on investigating the impact of omitted variables on the performance of several commonly used machine learning algorithms. Of course, in practice the data analyst may not realize that important variables have been omitted from the

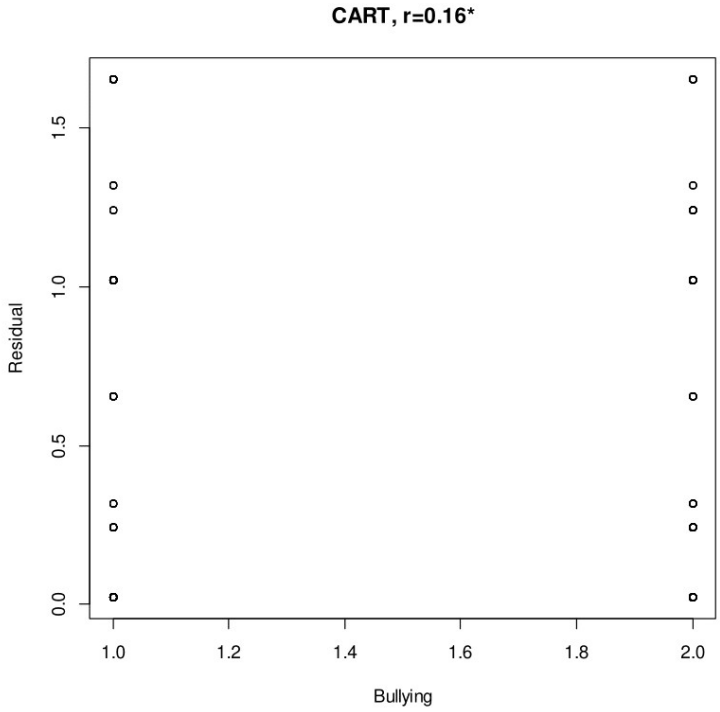
model. Therefore, it is important that the researcher have tools for ascertaining whether important variables have been omitted from the model. An approach for doing this that has been recommended in the regression literature (e.g., Fox, 2016; Pedhazur, 1997) involves plotting model residual values by individual independent variables that were used in the model. For a correctly specified model (i.e., a model with the important independent variables included) the residuals should not exhibit relationship with the individual independent variables that were included in the model. When the plot demonstrates a non-random relationship between the residuals and a predictor, the researcher would conclude that an important variable may have been omitted from the model. In addition, the correlation coefficient between the residuals and each of the independent variables can also be calculated for investigating the possibility of omitted variables being an issue. If the model is correctly specified (i.e., no important variables are omitted), these correlations should be near 0.

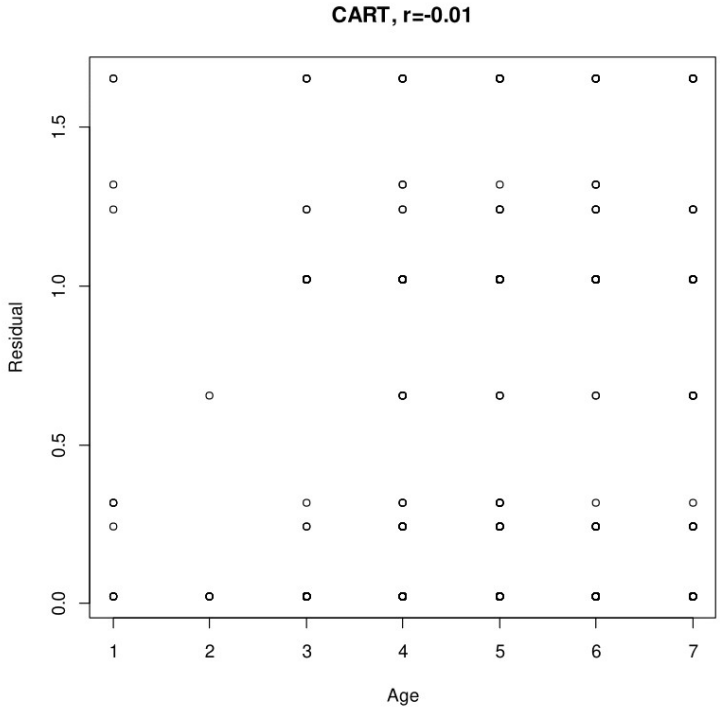
For the current example, the residuals were first plotted for the variable Bullied at School because it exhibited the greatest change in variable importance when the Considered Suicide variable was removed from the model. The plots for each of the three prediction methods studied here appear in Figure 4, with the biserial correlations appearing in the title of each graph. From these results, we can see that for both CART and RF, the relationship between the Bullying variable and the residuals was statistically significant and in the small range based on Cohen's (1988) guidelines. In contrast, for the Boosted trees model, the biserial correlation was not statistically significant and was negligible in size. Next, the residuals were plotted against participant age, as it did not exhibit a major shift in importance when Considered Suicide was removed from the model. The results presented in Figure 4 show that the correlations between age and the residuals were not statistically significant and were in the negligible for the three methods featured here. This result demonstrates that when an unimportant variable is missing there was not a relationship between age and the residuals. Taken together, the results for bullying and age demonstrate that use of the residual plots and correlation coefficients with CART and RF appear to be useful for identifying the presence of important missing variables in these two modeling techniques.

5 Discussion

The purpose of this study was to examine the impact on the relative importance measures of three popular machine learning algorithms, CART, RF, and Boosting, of omitting important predictor variables. The results presented above reveal that omitting such variables can indeed prove problematic, but that the algorithms are not impacted in the same way. As described above, when the correct model, including all relevant predictors, is fit to the data, the more important variable(s) are identified as such. In addition, it appears that the relative importance measures are somewhat sensitive to the role of interactions in determining which variables contribute most to the outcome. On the other hand, when a predictor that was associated with the outcome was omitted from the model, the relative importance values for the remaining variables in the model were clearly impacted. As would be expected, when fewer variables were in the model, the relative importance of all variables increased. This result would be expected, given that there are fewer variables among which the overall importance would need to be shared. More interestingly perhaps, was that a variable that was involved

Panel A: CART

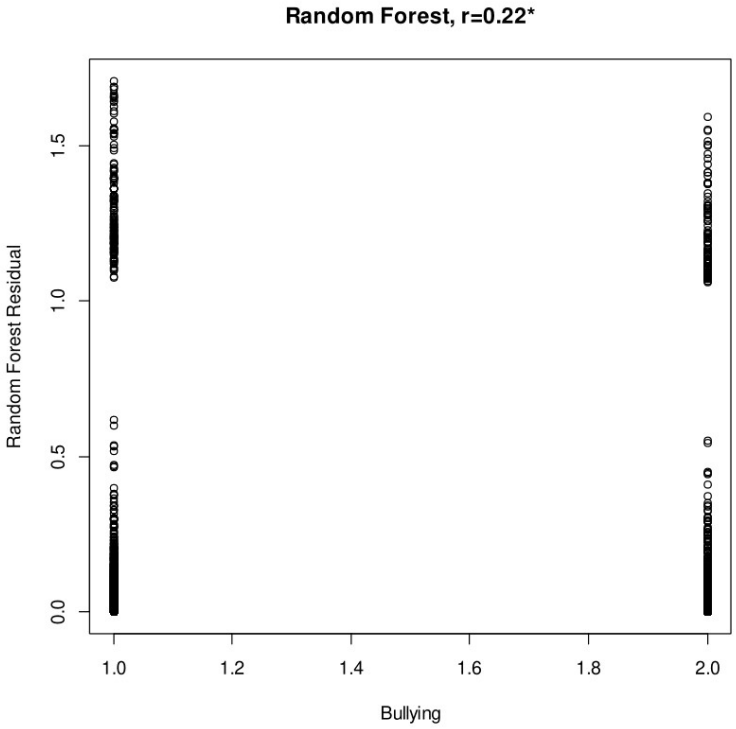


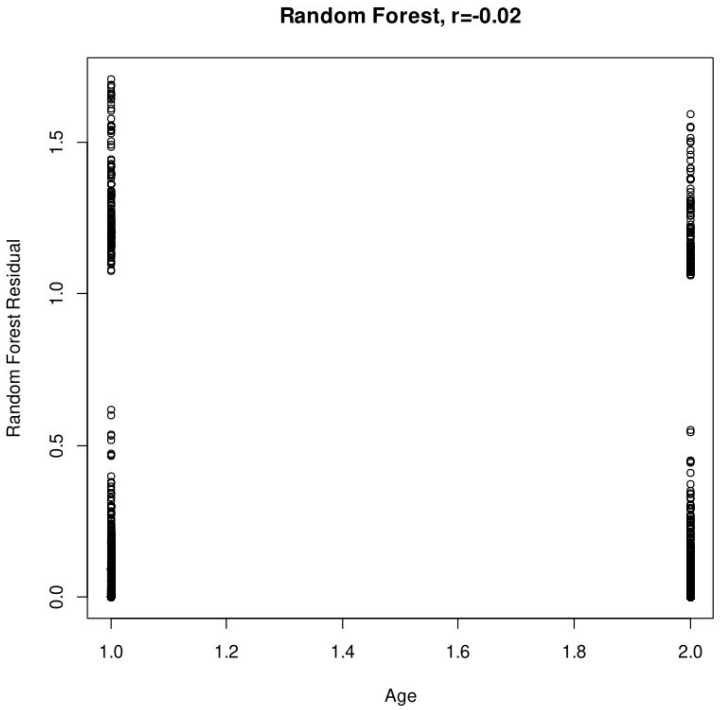


*Statistically significant for α=0.05

Figure 7: Residual by Bullying plots for CART, RF, and Boosted tree models *(continued on next page)*

Panel B: Random Forest

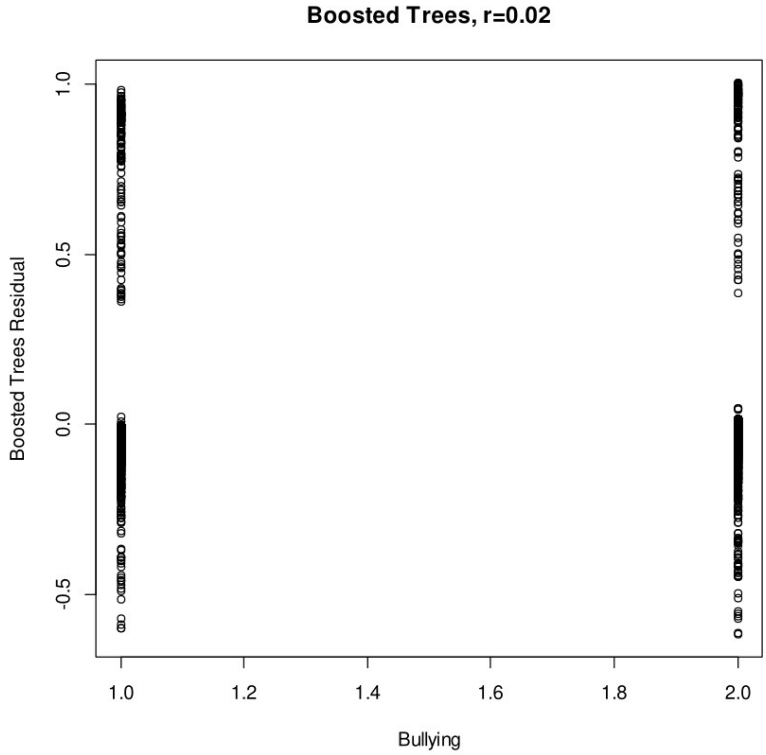




*Statistically significant for α=0.05

Figure 8: Residual by Bullying plots for CART, RF, and Boosted tree models *(continued on next page)*

Panel C: Boosted Trees



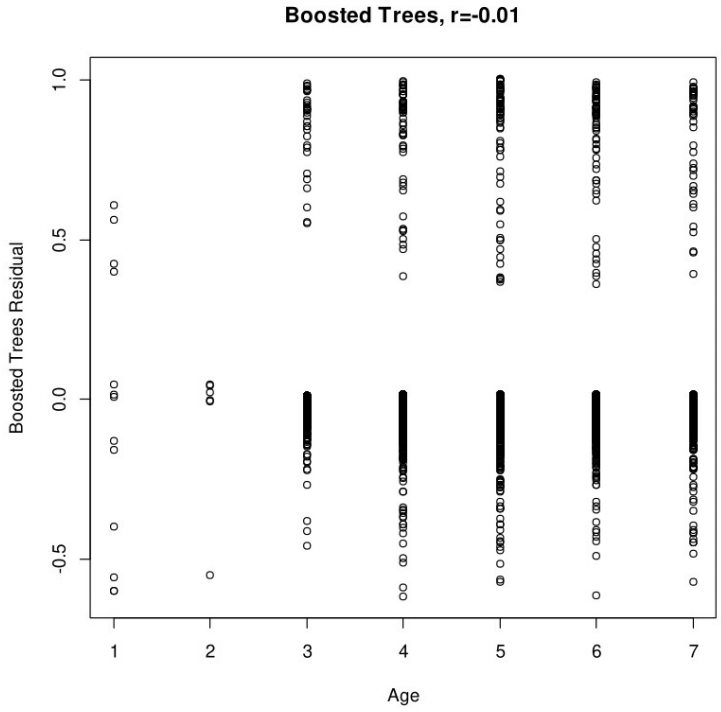


Figure 9: Residual by Bullying plots for CART, RF, and Boosted tree models

in an interaction with the omitted predictor showed a larger increase in relative importance than did one that was not associated with the omitted variable in any way. Thus, it would appear that the importance of the omitted predictor was, to some extent, transferred to the remaining variable with which it was related. This effect was magnified by a larger correlation with the omitted variable, by a stronger interaction effect with it, and by a stronger relationship between the omitted variable and the response. Finally, the results presented above, particularly those for the empirical example, suggest that the relative importance measure associated with CART is particularly sensitive to the omission of important variables from the prediction model. RF and boosting, while certainly impacted as well, were less affected in this study

The results of this study have a number of implications for statistical practice. Certainly first and foremost, they highlight the importance of researchers carefully considering which variables should be included in their study, and ensuring that these are measured, if at all possible. Excluding important variables will not only limit the utility of a model conceptually, but could also impact interpretations made by researchers regarding the importance of these variables. Indeed, even when the omitted predictor has only a moderate relationship with the outcome, its exclusion can have a dramatic impact on the relative importance values of the remaining variables in the model. This finding appears to be most important for CART, which not only exhibited changes in the absolute values of the relative important measures for individual variables, but also had a change in the ordering of the variables by importance. In other words, exclusion of an important model predictor could lead researchers to draw very incorrect conclusions about which variables are most important in terms of predicting a specific dependent variable. This leads to a second important implication for data analysts, which is that if a researcher's primary goal is to identify which variables are most strongly related to the outcome, and they are unsure whether all important variables have been included in the model, then they may be best off using RF or boosting rather than CART.

A third implication of this study is that when all of the relevant variables are included in the model, main effects appear to play a larger role in determining variable importance for CART, RF, and Boosting than do interactions. Figure 1 shows that only when the coefficient for the interaction of X_3 and X_4 is 0.8, their correlation is 0.8, and the coefficient for X_3 is 1.5 is the relative importance of X_4 comparable to that of X_5. Given that the main effect for X_5 was simulated to be 0.7, whereas that of X_4 was 0.2, it would appear that the relative importance measures studied here are more strongly influenced by the main effects than by the interactions.

A fourth implication, which can be drawn from the empirical example, is that the use of residual/independent variable plots and correlations may be an effective tool for identifying the presence of important missing variables with both CART and RF. In this example, there was are relationship between the residuals and the bullied in school variable for these models when the Considered Suicide variable was excluded, whereas for age no such relationship was found. Given that the removal of Considered Suicide did alter the variable importance values for bullying but not age, it appears that examining the relationships involving residuals and these independent variables for the CART and RF models may be helpful in identifying when important missing variables have been excluded. On the other hand, this approach did not yield similar results for boosted trees, meaning that it may not be as useful for that modeling strategy.

Directions for future research

The results of this study point to several directions for additional work. First, future simulation work should examine the impact of omitting relevant predictors on the relative importance measures when the outcome variable is binary. This study did include an empirical example with a binary outcome as a way of showing the potential impact of such omissions, but future work involving simulated data would be helpful in this regard. Future research should also examine the impact omitting non-normal variables from the model. In particular, it would be of interest to see the impact of omitting important nominal predictors with differing numbers of categories. A third area for future research involves the simulation of different model structures. The model selected here is based upon results from an existing dataset, but clearly other model structures would be of interest as well. Specifically, models including more predictor variables, a mix of categorical and continuous predictors, and more complex nonlinear terms would provide additional useful information for researchers using these prediction algorithms. Finally, and perhaps most importantly, future work should focus on approaches for ameliorating the impact of omitting important variables from the model. The results of the current study seem to indicate that more ensemble algorithms, such as RF and Boosting, which rely on multiple recursive partitioning trees, may yield more dependable results than was the case for a single tree produced by CART. However, it is also clear that more work needs to be done to determine under what conditions this is the case, and why it might be so. The current study is a first step in this direction. Finally, the possibility of using the relationships between residuals and independent variables for identifying when important predictors have been omitted from the model should be examined in future research with these data mining techniques. The empirical example appears to show that examining scatterplots and correlations between the residuals and individual independent variables may be quite helpful for this purpose, but a more systematic simulation study needs to be conducted in order to more fully understand this issue.

Conclusions

The results of this study build upon earlier work (Nystrom et al., 2019) by examining the impact of omitting important predictors from popular machine learning algorithms when the underlying model is nonlinear in nature. It seems clear that, just as for regression, the omission of important variables from a prediction model can result in major changes to the relative importance values for the remaining predictors in the model. The impact of the omitted variables was apparently transferred, at least in part, to the remaining predictor(s) with which it was associated. The result was that when the outcome variable was continuous in nature, the ordering of the predictors in terms of their importance might be altered, thus leading researchers to mistakenly attribute greater import to the variable than it warrants. Thus, researchers must be extremely vigilant when designing a study so that they include the relevant variables in their analyses. Not doing so could lead to mistaken conclusions regarding which variables are most associated with the outcome of interest. This result was not so strongly in evidence for RF or Boosting in the empirical example. However, given that the true model underlying that data is not known, those results must be given somewhat less weight than those associated with the simulation itself. It is hoped that these results provide

researchers with both a caveat for practice (carefully consider which variables are important and be sure to measure them if at all possible), and directions for future investigation.

References

Afshartous, D. and Preston, R. A. (2011), "Key results of interaction models with centering," *Journal of Statistics Education*, 19, 1–24.

Bauer, E. and Kohavi, R. (1999), "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine Learning*, 36, 105–139.

Bone, D., Bishop, S. L., Black, M. P., Goodwin, M. S., Lord, C., and Narayanan, S. S. (2016), "Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion," *Journal of Child Psychology and Psychiatry*, 57, 927–937.

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

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees (The Wadsworth Statistics/Probability Series)*, Chapman & Hall.

Christensen, K., Nørskov, S., Frederiksen, L., and Scholderer, J. (2017), "In search of new product ideas: Identifying ideas in online communities by machine learning and text mining," *Creativity and Innovation Management*, 26, 17–30.

Cramer, J. S. (2005), "Omitted variables and misspecified disturbances in the logit model," Tinbergen Institute Discussion Papers 05-084/4, Tinbergen Institute.

Delgadillo, J. and Gonzalez Salas Duhne, P. (2020), "Targeted prescription of cognitive–behavioral therapy versus person-centered counseling for depression using a machine learning approach." *Journal of Consulting and Clinical Psychology*, 88, 14–24.

Dias, R. D., Gupta, A., and Yule, S. J. (2019), "Using machine learning to assess physician competence: a systematic review," *Academic Medicine*, 94, 427–439.

Dietterich, T. G. (2000), "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning*, 40, 139–157.

Fox, J. (2016), *Applied regression analysis and generalized linear models*, Sage Publications.

Freund, Y. and Schapire, R. E. (1997), "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, 55, 119–139.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001), *The elements of statistical learning*, Springer series in statistics New York.

Gautier, N., Aider, J.-L., Duriez, T., Noack, B., Segond, M., and Abel, M. (2015), "Closed-loop separation control using machine learning," *Journal of Fluid Mechanics*, 770, 442–457.

Hew, K. F., Qiao, C., and Tang, Y. (2018), "Understanding student engagement in large-scale open online courses: A machine learning facilitated analysis of student's reflections in 18 highly rated MOOCs," *International Review of Research in Open and Distributed Learning*, 19, 69–93.

Hölters, O., Wert, B., Wietschel, M., Arens, M., Dötsch, C., Herkel, S., Krewitt, W., Markewitz, P., Möst, D., Scheufen, M., et al. (2002), "Econometric Analysis of Cross Section and Panel Data," .

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An introduction to statistical learning*, vol. 112, Springer.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., et al. (2005), *Applied linear statistical models*, McGraw-Hill New York.

Nystrom, E., Sharp, J. L., and Bridges, W. C. (2019), "The impact of correlated and/or interacting predictor omission on estimated regression coefficients in linear regression," *Journal of Statistical Theory and Practice*, 13, 56.

Pedhazur, E. (1997), *Multiple regression in behavioral research: Explanation and prediction*, New York: Wadsworth/Thomas Learning.

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., and Medina, D. (2018), "Effective injury forecasting in soccer with GPS training data and machine learning," *PloS one*, 13, e0201264.

Zhang, H. and Singer, B. H. (2010), *Recursive partitioning and applications*, Springer Science & Business Media.

Zhang, L. and Zhang, D. (2016), "Robust visual knowledge transfer via extreme learning machine-based domain adaptation," *IEEE Transactions on Image Processing*, 25, 4959–4973.

Zhu, Y., Zhou, L., Xie, C., Wang, G.-J., and Nguyen, T. V. (2019), "Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach," *International Journal of Production Economics*, 211, 22–33.

Received: *January 15, 2021*
Accepted: *November 3, 2021*