Journal of Data Science 19 (4), 569–592
October 2021
DOI: 10.6339/21-JDS1023
Statistical Data Science

Variable Importance Scores

Wei-Yin Loh[1,*] and Peigen Zhou[1]

[1]*Department of Statistics, University of Wisconsin, 1300 University Avenue, Madison, WI 53706, USA*

Abstract

There are many methods of scoring the importance of variables in prediction of a response but not much is known about their accuracy. This paper partially fills the gap by introducing a new method based on the GUIDE algorithm and comparing it with 11 existing methods. For data without missing values, eight methods are shown to give biased scores that are too high or too low, depending on the type of variables (ordinal, binary or nominal) and whether or not they are dependent on other variables, even when all of them are independent of the response. Among the remaining four methods, only GUIDE continues to give unbiased scores if there are missing data values. It does this with a self-calibrating bias-correction step that is applicable to data with and without missing values. GUIDE also provides threshold scores for differentiating important from unimportant variables with 95 and 99 percent confidence. Correlations of the scores to the predictive power of the methods are studied in three real data sets. For many methods, correlations with marginal predictive power are much higher than with conditional predictive power.

Keywords *bias correction; classification and regression tree; missing values; prediction*

1 Introduction

The question of how to quantify the relative importance of variables has intrigued researchers for years. While it was largely of academic interest early on, the question has attracted much interest in the last two decades, due to the availability of data with large numbers of variables and the desire to interpret "black box" machine learning models; see, e.g., Bring (1994), Bi (2012), and Wei et al. (2015). Ribeiro et al. (2016) (see also Lundberg and Lee, 2017) argue that interpretability of a model and its predictions is important in gaining a user's trust. Clearly, such trust can only be won if the interpretations themselves are not incorrect.

A well-known black-box model is random forest (RF, Breiman, 2001), which consists of hundreds of unpruned regression trees. It uses a permutation-based scheme to generate importance scores that has been widely copied. Some researchers have observed, however, that RF score orderings do not always agree with those based on traditional methods. For example, Bureau et al. (2005) used RF to identify single-nucleotide polymorphisms (SNPs) predictive of disease and found that while SNPs that are highly associated with disease (measured by Fisher's exact test) tend to have high RF scores, the two orderings do not match. Díaz-Uriarte and Alvarez de Andrés (2006) selected genes in microarray data by iteratively removing 20% of the genes with the lowest RF scores at each step. They found that this yielded a smaller set of genes than linear discriminant analysis, nearest neighbor and support vector machine methods, and that the RF results were more variable.

*Corresponding author. Email: loh@stat.wisc.edu.

Received July 6, 2021; Accepted August 26, 2021

Table 1: Variables in COVID data.

`died`	Died while hospitalized (0=no, 1=yes)
`agecat`	Age group (0=18–50, 1=50–59, 2=60–69, 3=70–79, 4=80–90 years)
`race`	White, Black or African American, Asian, Native Hawaiian or other Pacific Islander, American Indian or Alaska Native, Unknown
`sex`	Gender (male, female)
`aids`	AIDS/HIV (0=no, 1=yes)
`cancer`	Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin (0=no, 1=yes)
`cerebro`	Cerebrovascular disease (0=no, 1=yes)
`charlson`	Charlson comorbidity index (0, 1, . . . , 20)
`CHF`	Congestive heart failure (0=no, 1=yes)
`CPD`	Chronic pulmonary disease (0=no, 1=yes)
`dementia`	Dementia (0=no, 1=yes)
`diabetes`	Diabetes mellitus (0=no, 1=yes)
`hemipara`	Hemiplegia or paraplegia (0=no, 1=yes)
`metastatic`	Metastatic solid tumor (0=no, 1=yes)
`MI`	Myocardial infarction (0=no, 1=yes)
`mildliver`	Mild liver disease (0=no, 1=yes)
`modsevliv`	Moderate or severe liver disease (0=no, 1=yes)
`PUD`	Peptic ulcer disease (0=no, 1=yes)
`PVD`	Peripheral vascular disease (0=no, 1=yes)
`RD`	Rheumatic disease (0=no, 1=yes)
`renal`	Renal disease (0=no, 1=yes)

The differences in orderings may be demonstrated on a data set from Harrison et al. (2020) of 31,461 patients aged 18–90 years diagnosed with the COVID-19 disease between January 20 and May 26, 2020, in the United States. Table 1 lists the 21 variables, which consist of death during hospitalization, age group, sex, race, 16 comorbidities, and Charlson comorbidity index (risk score computed from the comorbidities). The authors estimated mortality risk by fitting a multiple linear logistic regression model, without Charlson index, to each age group. They found 10 variables statistically significant at the 0.05 level (without adjusting for multiplicity), namely, race, sex, and histories of myocardial infarction (`MI`), congestive heart failure (`CHF`), dementia, chronic pulmonary disease (`CPD`), mild liver disease (`mildliver`), moderate/severe liver disease (`modsevliv`), renal disease (`renal`), and metastatic solid tumor (`metastatic`).

Figure 1 shows the importance scores of the top 10 variables obtained from 12 methods discussed below. There is substantial variation in the orderings, although `agecat`, `charlson`, and `renal` are ranked in the top 3 by 7 of the 12 methods. Of the variables that Harrison et al. (2020) found statistically significant, `CPD` is not ranked in the top 10 by any method, and `mildliver` and `metastatic` are ranked in the top 10 only thrice and once, respectively. On the other hand, the non-significant variables `cancer`, `cerebro`, `diabetes`, `hemipara`, and `PVD` are ranked in the top 10 by 5, 10, 7, 3, and 9 methods, respectively. Thus statistical significance is not necessarily consistent with the importance scores.

What is one to do in the face of such disparate results? One solution is to average the ranks across the methods, but this assumes that the methods are equally good. Strobl et al. (2007),

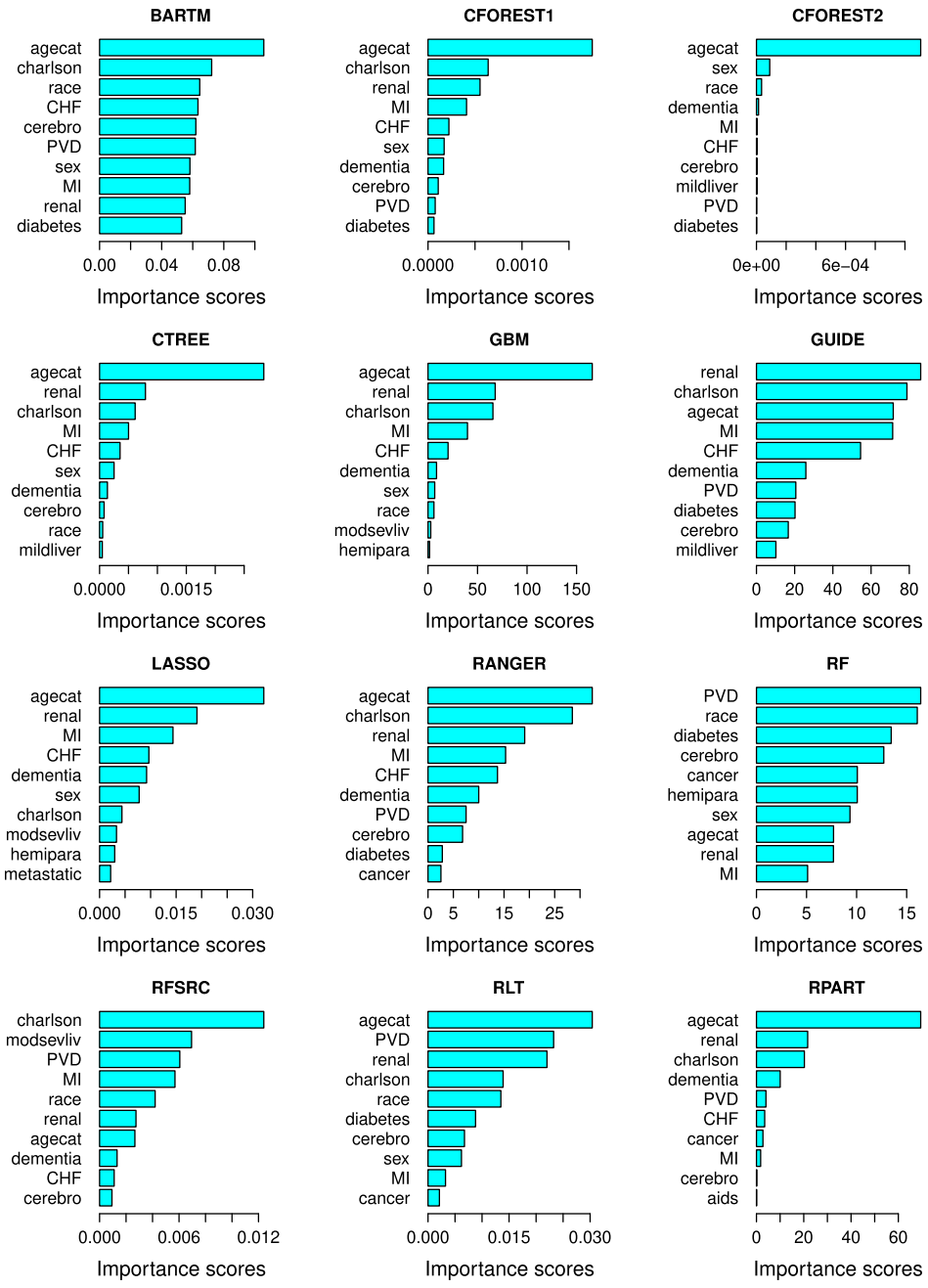

Figure 1: Top 10 variables for COVID data; LASSO, RANGER, RF, RFSRC, and RLT scores are averages over 100 trials with different random seeds.

Sandri and Zuccolotto (2008), and others have shown that RF scores are unreliable because they are biased towards certain variable types. A method is said to be "unbiased" if all predictor variables have the same mean importance score when they are independent of the response variable. One goal of this paper is to find out if there are other methods with such bias.

For a given data set, bias may be detected by estimating the mean scores over random permutations of the response variable, keeping the values of the predictor variables fixed. Let

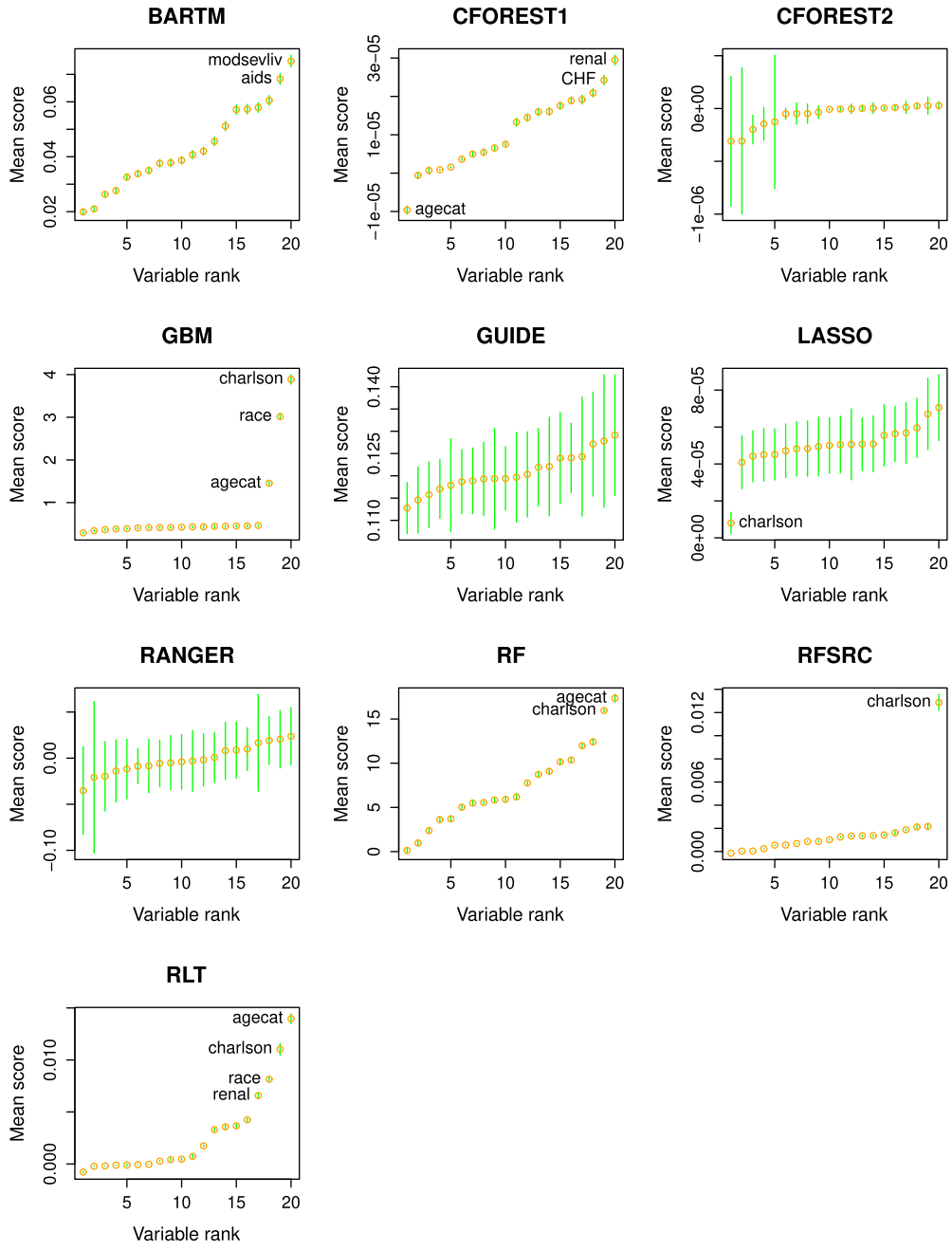

Figure 2: Mean importance scores $\overline{\texttt{VI}}$ and 2-SE bars from 1000 random permutations of the dependent variable for COVID data. Variables ordered by increasing mean scores. CTREE and RPART returned trees with no splits (hence no importance scores) for all permutations.

$\texttt{VI}_j(X)$ $(j = 1, 2, \ldots, J)$ denote the importance score of variable X in the jth permutation. Figure 2 plots the values of $\overline{\texttt{VI}}(X) = J^{-1} \sum_j \texttt{VI}_j(X)$ in increasing order and their 2-standard error bars, for $J = 1000$. An unbiased method should have all its error bars overlapping. The plots show that only CFOREST2, GUIDE, and RANGER have this property.

The remainder of this article is organized as follows. Section 2 describes the GUIDE method of calculating importance scores. Section 3 reviews the other 11 methods. Section 4 presents the results of simulations to show how bias affects their importance scores. Section 5 examines the extent to which the scores of each method are consistent with two measures of predictive power of the variables. One problem closely related to importance scoring is obtaining a threshold for distinguishing the important from unimportant variables. There have been few attempts at solving this problem, despite its being central to variable selection, particularly if the number of variables exceeds the sample size. Section 6 describes a general procedure incorporated in GUIDE for producing thresholds such that, with high probability, variables independent of the response will score below the thresholds. Section 7 shows how the GUIDE method applies to data with missing values and Section 8 concludes the article with some remarks.

2 GUIDE Scoring Method

The core GUIDE algorithm for constructing regression and classification trees is described in Loh (2002) and Loh (2009), respectively. It differs from CART (Breiman et al., 1984) in every respect except tree pruning, where both employ the same cost-complexity cross-validation technique. Whereas CART uses greedy search to select the split that most decreases node impurity, GUIDE uses a two-step procedure that (i) selects the split variable most highly associated with the response and (ii) then finds the best set of split values of the chosen variable to maximally decrease node impurity. This paradigm change started in Loh and Vanichsetakul (1988) and evolved principally through Chaudhuri et al. (1994), Loh and Shih (1997), and Loh (2002, 2009). Besides reducing computation, the technique saves GUIDE from biases in variable selection due to greedy search. Another important difference between GUIDE and CART is how each deals with missing values in predictor variables. CART uses surrogate splits to pass observations with missing values through each split, but this has been shown to be another source of selection bias (Kim and Loh, 2001). GUIDE, on the other hand, treats missing values as qualitative observations and sends them at each split to the node that yields more reduction in node impurity (Loh et al., 2019, 2020).

An earlier GUIDE importance scoring method was proposed in Loh (2012) but it has two weaknesses: (i) it is not unbiased if there is a mix of ordinal and categorical variables and (ii) it is not sensitive to local pairwise interactions. We now introduce an improved version that removes these deficiencies. As in the earlier method, it uses a weighted sum of chi-squared statistics obtained from a short (four-level) unpruned tree, but it adds conditional tests for pairwise interactions and a permutation-based step for bias correction. It is presented here for regression for simplicity, although it is applicable to classification as well.

Given a node t of a regression tree, let n_t denote the number of observations in t. The following steps are performed recursively, starting with the root node, until a tree with four levels of splits is obtained. Values of parameters m and d_k are chosen to control the probability of small cell counts in the contingency tables.

1. Fit a constant to the response values in t and compute the residuals.
2. Define a class variable Z such that $Z = 1$ if the observation has a positive residual and $Z = 2$ otherwise.
3. Define $m = 3$ if $n_t < 60$ and $m = 4$ otherwise.
4. For each ordinal variable X_k, let d_k denote the number of its distinct values (including missing value indicator) in t.

(a) If $d_k \leqslant 4$ or if $d_k = 5$ and X_k has missing values in t, define X'_k to be the categorical variable with nominal values being those of X_k (with a separate value for missing X_k values).
(b) Otherwise, define $q(0, k) = -\infty$.
   i. If X_k has missing values in t, let $q(i, k)$ be the sample $i/(m-1)$-quantile of the nonmissing values of X_k, for $i = 1, 2, \ldots, (m-2)$. Define categorical variable X'_k with nominal values

$$X'_k = \begin{cases} q(i,k), & \text{if } X_k \text{ is nonmissing and } q(i-1,k) < X_k \leqslant q(i,k) \\ q(m-2,k)+1, & \text{if } X_k \text{ is nonmissing and } X_k > q(m-2,k) \\ q(m-2,k)+2, & \text{if } X_k \text{ is missing} \end{cases}$$

   for $i = 1, 2, \ldots, (m-2)$.
   ii. If X_k has no missing values, let $q(i, k)$ be the sample i/m-quantile of the values of X_k, for $i = 1, 2, \ldots, (m-1)$. Define categorical variable

$$X'_k = \begin{cases} q(i,k), & \text{if } q(i-1,k) < X_k \leqslant q(i,k) \\ q(m-1,k)+1, & \text{if } X_k > q(m-1,k) \end{cases}$$

   for $i = 1, 2, \ldots, (m-1)$.
5. Define $X'_k = X_k$ for each categorical variable X_k.
6. For $k = 1, 2, \ldots, K$, where K is the number of variables, perform a contingency table chi-squared test of X'_k versus Z and denote its p-value by $p_1(k, t)$.
7. Initialize $j' = k' = 0$ and $p_2(j, k, t) = 1$ for $j, k = 1, 2, \ldots, K$. If $\min_k p_1(k, t) \geqslant 0.10/K$ (first Bonferroni correction), carry out the following interaction tests.
   (a) Transform each ordinal X_k to a 3-level categorical variable X'_k. If X_k has no missing values, X'_k is X_k discretized at the 33rd and 67th sample quantiles. If X_k has missing values, X'_k is X_k discretized at the sample median with missing values forming the third category. If X_k is a categorical variable, let $X'_k = X_k$.
   (b) For every pair (X'_j, X'_k) with $j < k$, perform a chi-squared test with the Z values as rows and the (X'_j, X'_k) values as columns and let $p_2(j, k, t)$ denote its p-value.
   (c) Let $(X'_{j'}, X'_{k'})$ be the pair of variables with the smallest value of $p_2(j, k, t)$. If $p_2(j', k', t) < 0.20\{K(K-1)\}^{-1}$ (second Bonferroni correction), set $p_1(j', t) = p_1(k', t) = p_2(j', k', t)$.
8. If $\min_{1 \leqslant j,k \leqslant K} p_2(j, k, t) < 0.20\{K(K-1)\}^{-1}$, split node t on either $X_{j'}$ or $X_{k'}$, whichever yields the larger decrease in node impurity (sum of squared residuals). Otherwise, split t on X_{k^*}, where k^* is the smallest value of k such that $p_1(k^*, t) = \min_k p_1(k, t)$.

Unadjusted importance score of X_k is given by

$$v(X_k) = \sum_t \sqrt{n_t}\, \chi^2_1(k, t) \tag{1}$$

where the sum is over the intermediate nodes and $\chi^2_1(k, t)$ denotes the $(1 - p_1(k, t))$-quantile of the chi-squared distribution with 1 degree of freedom.

The unadjusted scores $v(X_k)$ are slightly biased due partly to p-value differences between ordinal and categorical variables and partly to conditional step 7. To remove the bias, we standardize the scores by their expected values under the hypothesis that the response variable (Y) is independent of the X variables. Specifically, we randomly permute the Y values B times ($B = 300$ in the examples and simulations here) with the X values fixed and obtain the importance scores

from each permuted data set. Let $v_b^*(X_k)$ be the value of (1) in permutation $b = 1, 2, \ldots, B$, and define $\bar{v}(X_k) = B^{-1} \sum_b v_b^*(X_k)$. The GUIDE *bias-adjusted* variable importance score of X_k is

$$\texttt{VI}(X_k) = v(X_k)/\bar{v}(X_k). \tag{2}$$

3 Eleven Other Methods

This section briefly reviews 11 other importance scoring methods.

RPART. This is an R implementation of CART (Therneau and Atkinson, 2019b). Let $s = \{X \in A\}$ denote a split of node t for some variable X and set A, and let t_L and t_R denote its left and right child nodes. Given a node impurity function $i(t)$ at t, let $\Delta(s,t) = i(t) - i(t_L) - i(t_R)$ be a measure of the goodness of the split. For regression trees, $i(t) = \sum_{i \in t}(y_i - \bar{y}_t)^2$, where $\bar{y}_t$ is the sample mean at t. CART partitions the data with the split $s(t)$ that maximizes $\Delta(s,t)$. To evaluate the importance of the variables and to deal with missing values, CART finds the surrogate split $\tilde{s}_k(t)$ based on each X_k that best predicts $s(t)$. The importance score of X_k is measured by $\sum_t \Delta(\tilde{s}_k(t), t)$, with the sum over the intermediate nodes of the pruned tree (Breiman et al., 1984, pp. 141–147).

RPART measures importance differently from CART (Therneau and Atkinson, 2019a). Given a split $s(t)$ and a surrogate $\tilde{s}(t)$, let $j(s(t), \tilde{s}(t))$ be the total number of observations in t_L and t_R correctly sent by $\tilde{s}(t)$. Let n_L and n_R denote the numbers of observations in t_L and t_R, respectively. The "adjusted agreement" between s and $\tilde{s}$ is $a(s,\tilde{s}) = \{j(s,\tilde{s}) - \max(n_L, n_R)\}/\min(n_L, n_R)$. X_k is called a "primary" variable if it is in s and a "surrogate" variable if it is in $\tilde{s}$. Let $P(k)$ and $S(k)$ denote the sets of intermediate nodes where X_k is the primary and surrogate variable, respectively. RPART defines $\texttt{VI}(X_k) = \sum_{t \in P(k)} \Delta(s(t), t) + \sum_{t \in S(k)} a(s(t), \tilde{s}(t)) \Delta(\tilde{s}(t), t)$ as the importance score of X_k. The simulation results in Section 4 below show that the scores are biased, because maximizing the decrease in node impurity induces a bias towards selecting variables that allow more splits (White and Liu, 1994; Loh and Shih, 1997). Additionally, if there are missing values, selection of the surrogate variables is biased too (Kim and Loh, 2001).

GBM. This is gradient boosting machine (Friedman, 2001). It uses functional gradient descent to build an ensemble of short CART trees. For a single tree, the importance score of a variable is the square root of the total decrease in node impurity (squared error in the case of regression) over the nodes where the variable appears in the split. For an ensemble, it is the root mean squared importance score of the variable over the trees (Friedman, 2001, p. 1217). We use the R function `gbm` (Greenwell et al., 2019) to construct the GBM models and the `varImp` function in the `caret` package (Kuhn, 2020) to calculate the importance scores.

RF. This is an R implementation of random forest (Liaw and Wiener, 2002). It has two measures for computing importance scores. The first is decrease in accuracy of the forest in predicting the "out-of-bag" (OOB) data before and after random permutation of the predictor variable, where the OOB data are the observations not in the bootstrap sample. The second uses decrease in node impurity, which is the average of the total decrease in node impurity of the trees. Partly due to CART's split selection bias, the second measure is known to be unreliable (Strobl et al., 2007; Sandri and Zuccolotto, 2008). The results reported here use the first measure.

RANGER. Sandri and Zuccolotto (2008) used pseudovariables to correct the bias in RF's decrease in node impurity measure. (Pseudovariables were employed earlier by Wu et al., 2007.) Given K predictor variables $\mathbf{X} = (X_1, X_2, \ldots, X_K)$, another K pseudovariables $\mathbf{Z} =$

$(Z_1, Z_2, \ldots, Z_K)$ are added where the rows of $\mathbf{Z}$ are random permutations of the rows of $\mathbf{X}$. The RF algorithm is applied to the $2K$ predictors and the importance score of X_k is adjusted by subtracting the score of Z_k for $k = 1, 2, \ldots, K$. This approach requires more computer memory and increases computation time (a forest has to be constructed for each generation of $\mathbf{Z}$). To partially circumvent this, Nembrini et al. (2018) proposed using only a single generation of $\mathbf{Z}$ and storing only the permutation indices rather than the values of $\mathbf{Z}$. Their method is implemented in the `ranger` R package (Wright and Ziegler, 2017). The cost of using a single generation of $\mathbf{Z}$ is that the RANGER scores are even more random than the original RF scores which are themselves random (unless the random seed is locked). As a result, there are no savings in computation time in real applications because RANGER must be applied multiple times to stabilize the average importance scores. In the examples here, the RANGER scores are averaged over 100 replications.

RFSRC. This is another ensemble method based on RF (Ishwaran, 2007; Ishwaran et al., 2008). The importance of variable X is the difference between the prediction error of the OOB sample before and after X is "noised up". "Noising up" here means that if an OOB observation encounters a split on X at a node t, it is randomly sent to the left or right branch, with equal probability, *at* t *and all its descendent nodes*. Missing values in a predictor variable are imputed nodewise, by replacing each missing value with a randomly selected non-missing value in the node. The results for RFSRC here are obtained with the `randomForestSRC` R package (Ishwaran and Kogalur, 2007).

RLT. This method may be thought of as "RF within RF." Called "reinforcement learning trees" (Zhu et al., 2015), it constructs an ensemble of trees from bootstrap samples, but uses the RF permutation-based importance scoring method to select the most important variable to split each node in each tree. After the ensemble is constructed, the final importance scores are obtained once more using the RF permutation scheme. The results here are produced by the `RLT` R package (Zhu, 2018).

CTREE. This is the "conditional inference tree" algorithm of Hothorn et al. (2006). It follows the GUIDE approach of using significance tests to select a variable to split each node of a tree. But unlike GUIDE, CTREE uses linear statistics based on a permutation test framework and, instead of pruning, it uses Bonferroni-type p-value thresholds to determine tree size. Further, each significance test employs only observations with non-missing values in the X variable being evaluated. Observations with missing values are passed through each split by means of surrogate splits as in CART. Importance scores are obtained as in RFSRC, except that an OOB observation missing the split value at a node is randomly sent to the left or right child node with probabilities proportional to the samples sizes of the non-missing observations in the two child nodes. The results here are obtained with the `partykit` R package.

CFOREST. This is an ensemble of CTREE trees from the `partykit` package. Instead of bootstrap samples, it takes random subsamples (without replacement) of about two-thirds of the data to construct each tree. Strobl et al. (2007) showed that this removes a bias in RF that gives higher scores to categorical variables with large numbers of categories. It is the default option in `partykit`, which we denote by CFOREST1. Another option, which we denote by CFOREST2, is conditional permutation of the variables, which Strobl et al. (2008) recommended for reducing the bias in RF towards correlated variables.

LASSO. This is linear regression with the lasso penalty. The importance score of an ordinal variable is the absolute value of its coefficient in the fitted model and that of a categorical

Table 2: Simulation models $Y = \mu(X) + \epsilon$, with ϵ independent standard normal.

	Model	Expected highest scoring variables
E0	$\mu(X) = 0$	None; equal expected scores
E1	$\mu(X) = 0.2N_2$	N_2 followed by N_3 and N_4
E2	$\mu(X) = 0.1(N_1 + N_2)$	N_1 and N_2, followed by N_3 and N_4
E3	$\mu(X) = 0.2B_1$	B_1
E4	$\mu(X) = 0.2B_2$ $(= 0.2I(C_2 \leqslant 5))$	B_2 and C_2
E5	$\mu(X) = 0.5\{I(B_1 = 0, C_1 \leqslant 5) + I(B_1 = 1, C_1 > 5)\}$	B_1 and C_1

variable is the average of the absolute values of the coefficients of its dummy variables. All variables (including dummy variables) are standardized to have mean 0 and variance 1 prior to model fitting. We use the implementation in the `glmnet` R package (Friedman et al., 2010).

BARTM. This is `bartMachine` (Bleich et al., 2014), a Bayesian method of constructing a forest of regression trees based on the BART algorithm (Chipman et al., 2010). It models the response variable as a sum of regression tree models plus homoscedastic Gaussian noise. Prior distributions must be specified for all unknown parameters, including the set of tree structures, terminal node parameters, and the Gaussian noise variance. Following Bleich et al. (2014), the importance of a variable is the relative frequency that it appears in the splits in the trees. The results here are obtained from the `bartMachine` R package with default parameters.

4 Simulation Experiments

We performed 6 simulation experiments (E0–E5) involving 11 predictor variables (B_1, B_2, C_1, C_2, N_1, N_2, N_3, N_4, S_1, S_2, S_3) to compare the performance of the 12 methods. Variable sets $\{B_1\}$, $\{C_1\}$, $\{B_2, C_2\}$, $\{N_1, N_2, N_3, N_4\}$, and $\{S_1, S_2, S_3\}$ are mutually independent. Variable B_1 is Bernoulli with $P(B_1 = 1) = 0.50$, and C_1, C_2 are independent categorical variables taking values $1, 2, \ldots, 10$ with equal probability 0.10. Variable $B_2 = I(C_2 \leqslant 5)$ is a binary variable derived from C_2. Variable N_1 is independent standard normal and (N_2, N_3, N_4) is multivariate normal with zero mean, unit variance, and constant correlation 0.90. The triple (S_1, S_2, S_3) is obtained by setting $S_1 = \min(U_1, U_2)$, $S_2 = |U_1 - U_2|$, and $S_3 = 1 - \max(U_1, U_2)$, where U_1 and U_2 are independent and uniformly distributed variables on the unit interval, so that $S_1 + S_2 + S_3 = 1$ and $\text{cor}(S_i, S_j) = -0.50$ $(i \neq j)$.

Table 2 shows the models used to generate the dependent variable $Y = \mu(X) + \epsilon$, where $\mu(X)$ is a function of the predictor variables and ϵ an independent standard normal variable. The purpose of null model E0, where Y is independent of the X variables, is identification of methods with biased importance scores. The other models have one or two important variables each; they show how bias suppresses the scores of these variables. For each model, importance scores are obtained from 1000 simulation trials, with random samples of 400 observations per trial.

Figure 3 shows the average scores and their 2-SE (simulation standard error) bars for model E0. The 2-SE bars should overlap if there is no selection bias. We see that only CFOREST2, CTREE, GUIDE, and RANGER have this property. The RANGER scores of the binary variables

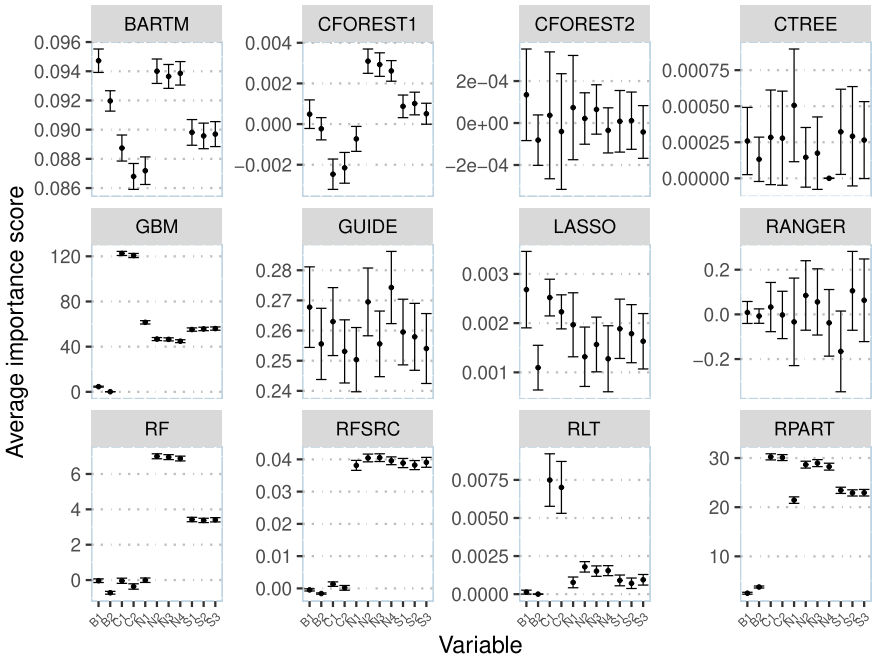

Figure 3: Average importance scores with 2-SE error bars for model E0, where predictor variables are independent of Y.

B_1 and B_2 are much less variable than those of the other variables. BARTM, CFOREST1 and RF are biased towards correlated variables N_2, N_3 and N_4. BARTM is also biased towards binary variables. GBM, RLT and RPART are biased towards multi-category variables C_1 and C_2. RFSRC is biased against all categorical variables and LASSO is biased *towards* B_1 but *against* B_2.

Figures 4–8 show boxplots of the simulated scores for models E1–E5, with darker color indicating important variables. The results may be summarized as follows.

E1. The model is $\mu(X) = 0.2N_2$, but the response is also associated with N_3 and N_4 through their correlation with N_2. Therefore these variables should have the three highest expected scores. Figure 4 shows that all but one method give their highest median scores to these three predictors. The exception is GBM, whose bias towards variables C_1 and C_2 frequently makes them appear more important than N_2, N_3 and N_4.

E2. The model is $\mu(X) = 0.1(N_1 + N_2)$, where N_1 is independent of N_2 but the latter is highly correlated with N_3 and N_4. We expect N_1 and N_2 to have equal and highest expected scores, with those of N_3 and N_4 close behind. Figure 5 shows that this is true of all methods except GBM, RF, RLT, and RPART. For RF and RPART, the presence of N_3 and N_4 raises the median score of N_2 above that of N_1. GBM again tends to incorrectly score C_1 and C_2 highest.

E3. The model is $\mu(X) = 0.2B_1$. Because B_1 is independent of the other predictors, we expect it to have the highest median scores. All except CTREE, GBM, LASSO, RF, RFSRC, and RPART show this. CTREE and LASSO fail because they have median scores of 0 for all variables. GBM and RPART fail due to bias towards C_1 and C_2. RF fails due to the high correlation among N_2, N_3 and N_4. RFSRC fails due to bias towards continuous variables.

Table 3: Causes of inaccuracies in scoring methods for Models E0–E5. (a) biased towards correlated variables N_2, N_3, N_4 and against independent variable N_1; (b) biased against non-dichotomous categorical variables C_1, C_2; (c) biased towards non-dichotomous categorical variables C_1 and C_2; (d) biased against dependent categorical variable C_2; (e) biased towards dichotomous variables B_1, B_2; (f) biased against dichotomous variables B_1, B_2; (g) biased towards B_1 and against B_2; (h) biased against all categorical variables; (z) median scores of zero for all variables.

Method	E0	E1	E2	E3	E4	E5
BARTM	*a,e*				*d,e*	
CFOREST1	*a,b*				*d*	
CFOREST2						
CTREE			*z*	*z*	*z*	*z*
GBM	*c,f*	*c*	*c*	*c,f*	*c,f*	*c,f*
GUIDE						
LASSO	*g*			*z*	*z*	*z*
RANGER						
RF	*a,h*		*a*	*a*		*a*
RFSRC	*h*			*h*	*h*	*h*
RLT	*c,f*		*c*	*c,f*	*c,f*	*c,f*
RPART	*c,f*		*c*	*c,f*	*c,f*	*c,f*

E4. The model is $\mu(X) = 0.2B_2$ but because $B_2 = I(C_2 \leqslant 5)$, variables B_2 and C_2 should have the highest expected scores. Only GUIDE, RANGER and possibly CFOREST2 have this property. BARTM and CFOREST1 give the highest median score to B_2 but middling median scores to C_2. GBM and RLT, due to their strong bias towards high-level categorical variables (Figure 3), give highest median score to C_2 but low or middling median score to B_2. As in model E3, CTREE and LASSO cannot reliably identify B_2 or C_2 as important because they have median scores of 0 for all predictors.

E5. The model is $\mu(X) = 0.5\{I(B_1 = 0, C_1 \leqslant 5) + I(B_1 = 1, C_1 > 5)\}$, which has an interaction between B_1 and C_1. BARTM, CFOREST1, CFOREST2, GUIDE, and RANGER correctly give these two predictors the highest median scores. On the other hand, GBM and RPART give B_1 the lowest median score due to their bias against binary variables. RF ranks with high frequency the correlated variables N_2, N_3, N_4 as most important. RFSRC also gives B_1 and C_1 low scores due to its bias against all categorical variables. RLT gives C_1 and C_2 the highest median scores due to its bias towards these two variables. As in E3 and E4, CTREE and LASSO are ineffective because both have median scores of 0 for all variables.

Table 3 lists the main reasons why some methods fail to correctly identify the important variables across the models. GBM, RFSRC, RLT, and RPART are highly unreliable if categorical variables are present. RF is highly unreliable if there are categorical variables or correlated continuous variable. CTREE and LASSO are often useless due to their high frequency of producing scores of 0 for all variables. CFOREST2, GUIDE, and RANGER are the only unbiased methods and therefore the most likely to correctly identify important variables.

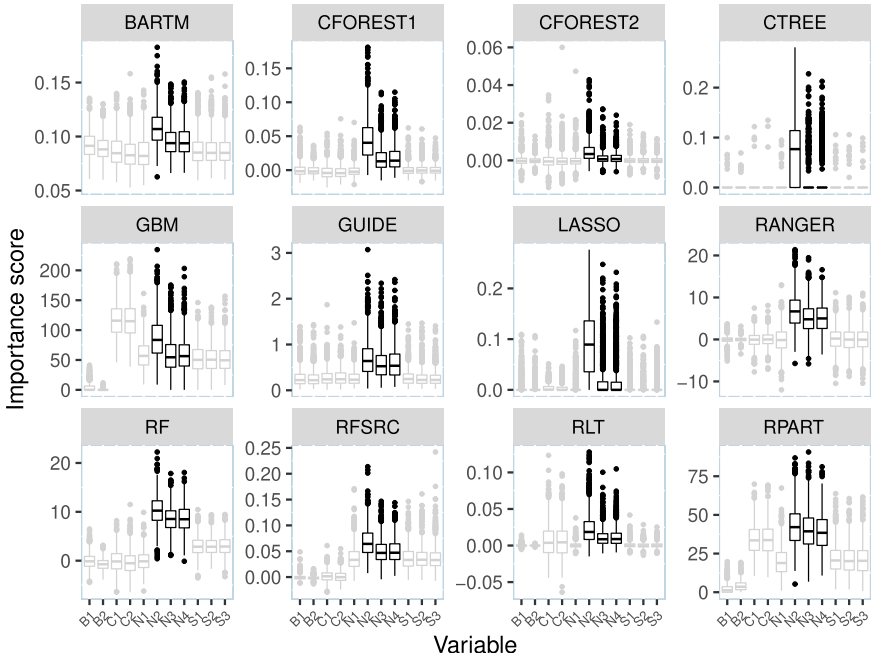

Figure 4: Boxplots of importance scores over 1000 trials for model E1, where $\mu(X) = 0.2N_2$ and N_2, N_3, N_4 are highly correlated. Important variables are in darker color.

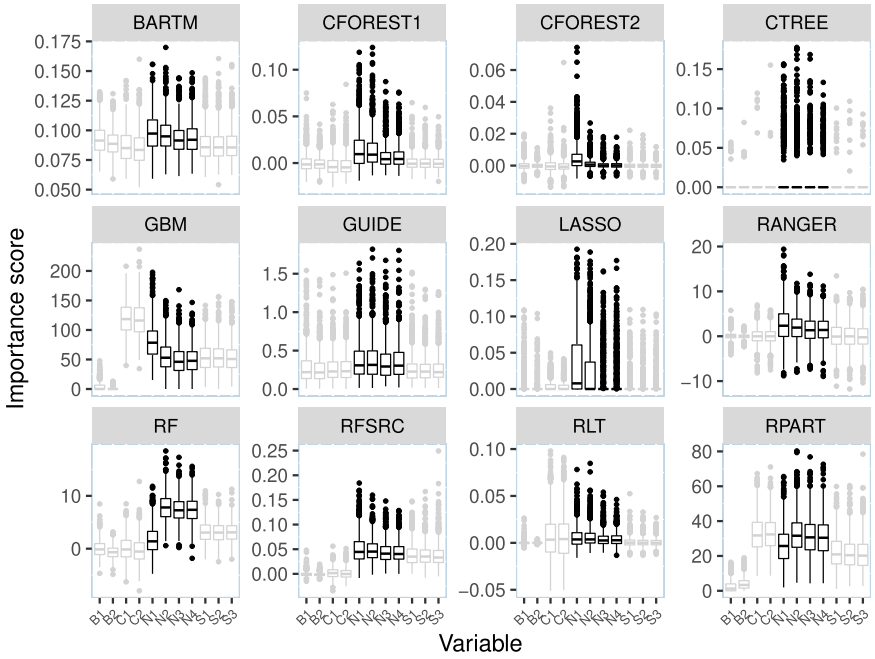

Figure 5: Boxplots of importance scores over 1000 trials for model E2 where $\mu(X) = 0.1(N_1 + N_2)$ and N_2, N_3, N_4 are highly correlated. Important variables are in darker color.

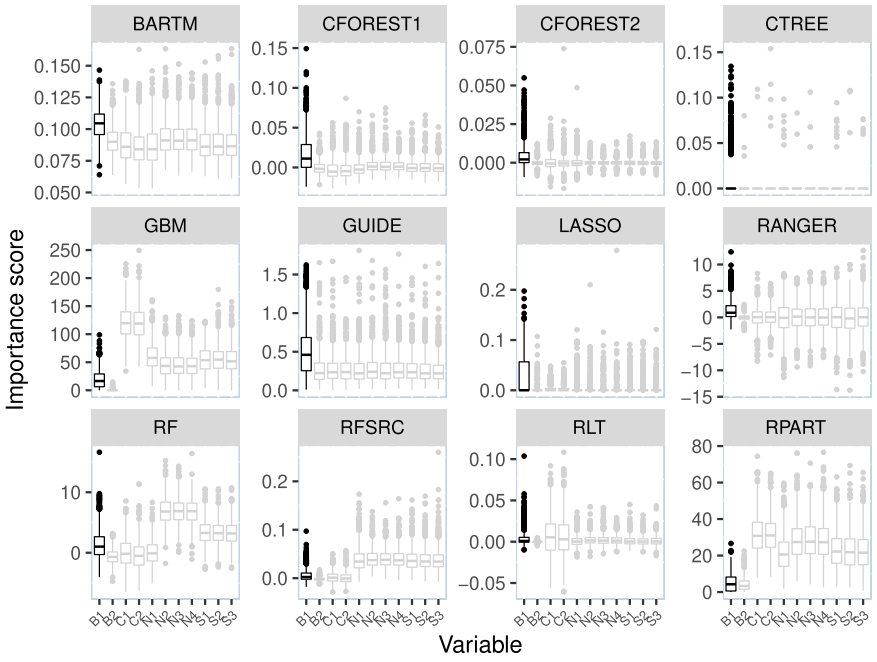

Figure 6: Boxplots of importance scores over 1000 trials for model E3, where $\mu(X) = 0.2B_1$. Important variables are in black.

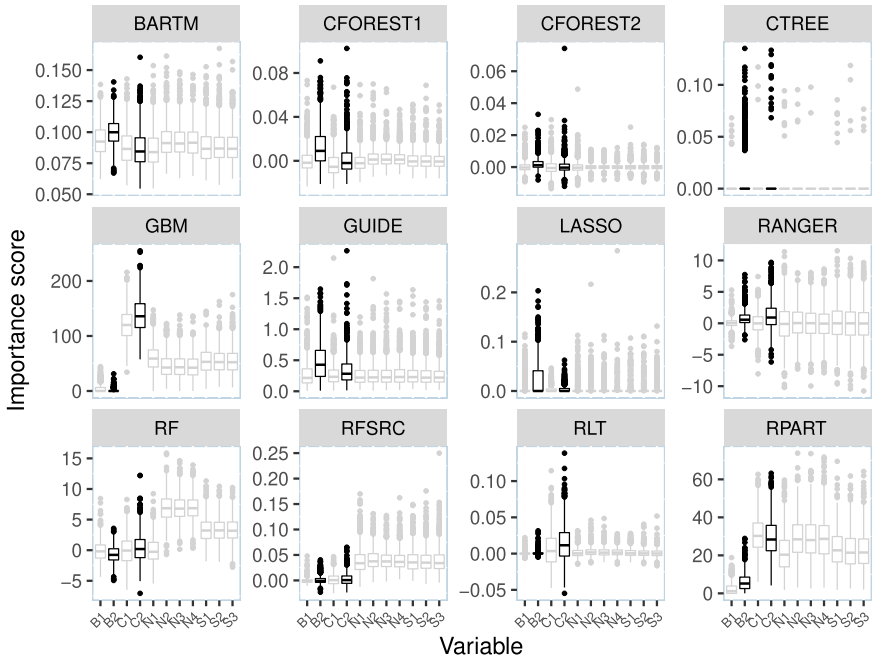

Figure 7: Boxplots of importance scores over 1000 trials for model E4, where $\mu(X) = 0.2B_2$ and $B_2 = I(C_2 \leqslant 5)$. Important variables are in darker color.

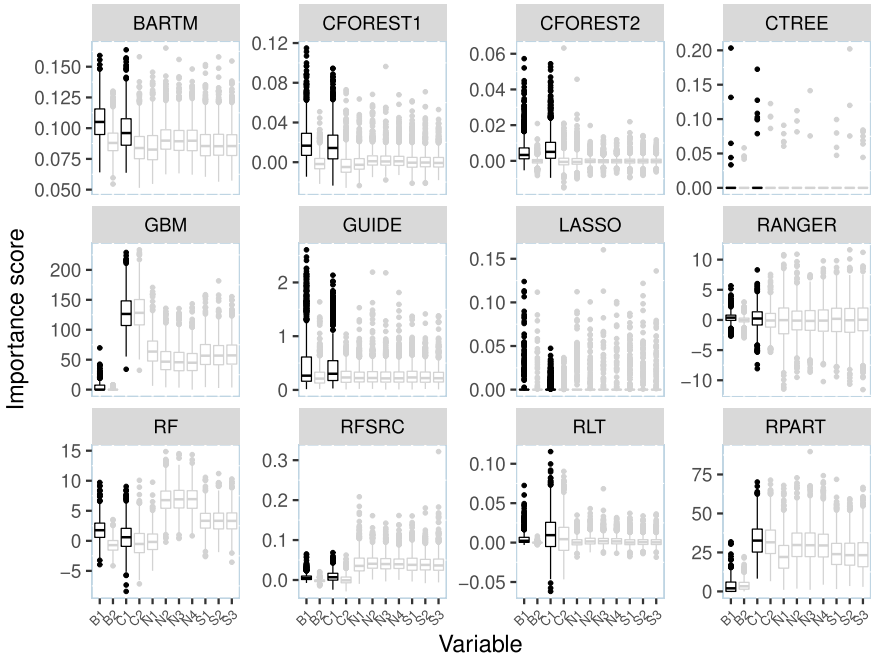

Figure 8: Boxplots of importance scores over 1000 trials for model E5, where $\mu(X) = 0.5\{I(B_1 = 0, C_1 \leqslant 5) + I(B_1 = 1, C_1 > 5)\}$. Important variables are in darker color.

5 Predictive Importance

"Predictive importance" may be interpreted as the effect of a variable on the prediction of a response, but it is not known which methods produce scores that reflect the concept. BARTM scores variables by their frequencies of being chosen to split the nodes of the trees. GBM and RPART base their scores on decrease in node impurity. LASSO uses absolute values of estimated regression coefficients. CFOREST, CTREE, RANGER, RF, and RFSRC measure change in prediction accuracy after random permutation of the variables—an approach that Strobl et al. (2008) call "permutation importance." Being based on chi-squared tests of association with the response variable at the nodes of a tree, GUIDE scores are measures of "associative importance."

To see how well scores reflect predictive importance, a precise definition of the latter is needed. Given predictor variables $X_1, X_2, \ldots, X_K$, consider the four models,

$$Y = \mu + \epsilon \tag{3}$$
$$Y = f_j(X_j) + \epsilon \tag{4}$$
$$Y = g_j(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_K) + \epsilon \tag{5}$$
$$Y = h(X_1, X_2, \ldots, X_K) + \epsilon \tag{6}$$

where μ is an unknown constant, f_j, g_j, and h are unknown functions of their arguments, and ϵ is an independent variable with zero mean and variance possibly dependent on the values of the X variables. Equation (3) states that Y is independent of the predictors, (4) states that Y depends only on X_j, (5) states that Y depends on all variables except X_j, and (6) allows Y to depend on all variables. Let $\hat{\mu}$, $\hat{f}_j$, $\hat{g}_j$, and $\hat{h}$ denote estimates of μ, f_j, g_j, and h, respectively,

obtained from a training sample. Define

$$\begin{aligned} S_0 &= E(Y - \hat{\mu})^2 \\ S_j &= E(Y - \hat{f}_j(X_j))^2 \\ S_{-j} &= E(Y - \hat{g}_j(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_K))^2 \\ S &= E(Y - \hat{h}(X_1, \ldots, X_K))^2 \end{aligned}$$

where the expectations are computed with $\hat{\mu}$, $\hat{f}_j$, $\hat{g}_j$, and $\hat{h}$ fixed. Call $(S_0 - S_j)$ the *marginal predictive value* of X_j because it is the difference in mean squared error between predicting Y with and without X_j, ignoring the other predictors. Call $(S_{-j} - S)$ the *conditional predictive value* of X_j because it is the difference in mean squared error between predicting Y without and with X_j, with the other predictors included.

Correlations between the importance scores and marginal and conditional predictive values indicate how well the former reflects the latter. To compute the correlations for a given data set, we need first to estimate μ, f_j, g_j, and h. Here we use the average of 5 ensemble methods, namely, CFOREST, GBM, GUIDE forest, RANGER, and RFSRC to do so. This ensures that no scoring method has an unfair advantage. We use leave-one-out cross-validation to estimate S_0, S_j, S_{-j}, and S. Specifically, given a data set $\{(y_i, x_{i1}, \ldots, x_{iK}), i = 1, 2, \ldots, n\}$, define the vectors and matrices

$$\begin{aligned} \mathbf{x}_j &= (x_{1j}, x_{2j}, \ldots, x_{nj})' \\ \mathbf{x}_j^{(-i)} &= (x_{1j}, x_{2j}, \ldots, x_{i-1,j}, x_{i+1,j}, \ldots, x_{nj})' \\ \mathbf{X} &= (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K) \\ \mathbf{X}^{(-i)} &= (\mathbf{x}_1^{(-i)}, \mathbf{x}_2^{(-i)}, \ldots, \mathbf{x}_K^{(-i)}) \\ \mathbf{X}_{(-j)} &= (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_K) \\ \mathbf{X}_{(-j)}^{(-i)} &= (\mathbf{x}_1^{(-i)}, \mathbf{x}_2^{(-i)}, \ldots, \mathbf{x}_{j-1}^{(-i)}, \mathbf{x}_{j+1}^{(-i)}, \ldots, \mathbf{x}_K^{(-i)}) \end{aligned}$$

where $(\mathbf{x}_j^{(-i)}, \mathbf{X}^{(-i)}, \mathbf{X}_{(-j)}^{(-i)})$ is $(\mathbf{x}_j, \mathbf{X}, \mathbf{X}_{(-j)})$ without the ith row and $(\mathbf{X}_{(-j)}, \mathbf{X}_{(-j)}^{(-i)})$ is $(\mathbf{X}, \mathbf{X}^{(-i)})$ without the jth column. Let $(\hat{f}_j^{(-i)}, \hat{g}_j^{(-i)}, \hat{h}^{(-i)})$ denote the function estimates of (f_j, g_j, h) based on $(\mathbf{x}_j^{(-i)}, \mathbf{X}_{(-j)}^{(-i)}, \mathbf{X}^{(-i)})$, respectively, obtained from the average of the 5 ensemble methods. Let $\bar{y} = n^{-1} \sum_k y_k$, $\bar{y}^{(-i)} = (n-1)^{-1} \sum_{k \neq i} y_k$ and define the leave-one-out mean squared errors

$$\begin{aligned} \hat{S}_0 &= n^{-1} \sum_{i=1}^{n} (y_i - \bar{y}^{(-i)})^2 \\ \hat{S}_j &= n^{-1} \sum_{i=1}^{n} \{y_i - \hat{f}_j^{(-i)}(x_{ij})\}^2 \\ \hat{S}_{-j} &= n^{-1} \sum_{i=1}^{n} \{y_i - \hat{g}_j^{(-i)}(x_{i1}, x_{i2}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{iK})\}^2 \\ \hat{S} &= n^{-1} \sum_{i=1}^{n} \{y_i - \hat{h}^{(-i)}(x_{i1}, x_{i2}, \ldots, x_{iK})\}^2. \end{aligned}$$

Denote the estimated marginal and conditional predictive values by $\texttt{MPV}_j = \hat{S}_0 - \hat{S}_j$ and $\texttt{CPV}_j = \hat{S}_{-j} - \hat{S}$, respectively. We compare them on the following three real data sets.

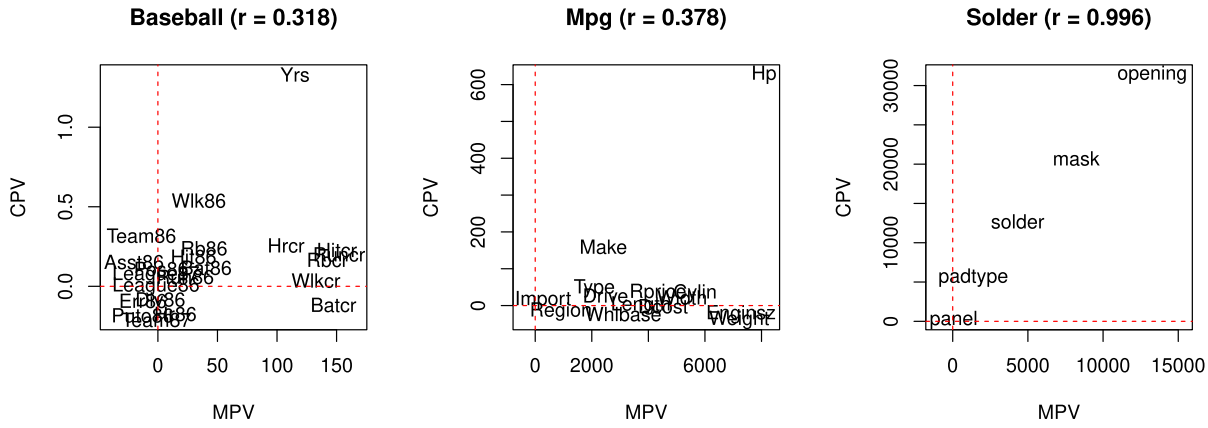

Figure 9: `CPV` versus `MPV` and their correlations for three data sets.

Baseball. The data give performance and salary information of 263 North American Major League Baseball players during the 1986 season (Denby, 1986). The response variable is log-salary and there are 22 predictor variables; see Hoaglin and Velleman (1995) and references therein for definitions of the variables. The plot on the left side of Figure 9 shows a rather weak correlation of 0.318 between `CPV` and `MPV`. Variable `Yrs` (number of years in the major leagues) has high values of `MPV` and `CPV` but `Batcr` (number of times at bat during career) has a high value of `MPV` and a negative value of `CPV`. This suggests that `Batcr` is an excellent predictor if it is used alone, but its addition after the other variables are included does not increase accuracy.

Mpg. This data set gives the characteristics, price, and dealer cost of 428 new model year 2004 cars and trucks (Johnson, 2004). We use 14 variables to predict city miles per gallon (mpg). The middle panel of Figure 9 shows that `Hp` (horsepower) has the highest values of `MPV` and `CPV`. Variable `Make` (which has 38 categorical values) has the second highest `CPV` but its `MPV` is below average, indicating that its predictive power is mainly derived from interactions with other variables. The correlation between `CPV` and `MPV` is 0.378.

Solder. Chambers and Hastie (1992) use data from a circuit board soldering experiment to demonstrate Poisson regression in R. The data, named `solder.balance` in the `rpart` R package, give the number of solder skips in an unreplicated $3 \times 2 \times 4 \times 10 \times 3$ factorial experiment. Because not all scoring methods are applicable to Poisson regression, we use least squares with square root of number of solder skips as dependent variable. The right panel of Figure 9 shows that, due to the factorial design, `CPV` and `MPV` are almost perfectly correlated.

Table 4 gives the correlations between the importance scores `VI` and `MPV` and `CPV` for each method and Figure 10 shows them graphically. They yield the following observations.

Baseball. The importance scores are highly correlated with `MPV` for GUIDE and RANGER, but not for LASSO where there is barely any correlation. On the other hand, the scores are weakly correlated with `CPV` for all methods except BARTM and LASSO. This may be due to many variables being correlated here.

Mpg. GUIDE and RANGER are again the two methods with importance scores most highly correlated with `MPV`; the correlations for the other methods range from 0.54 for RF to 0.85 for BARTM and RPART. For `CPV`, GBM has the highest correlation of 0.88, followed by

Table 4: Correlations between importance scores `VI` and marginal and conditional predictive values `MPV` and `CPV`.

	Baseball		Mpg		Solder	
Method	`MPV`	`CPV`	`MPV`	`CPV`	`MPV`	`CPV`
BARTM	0.75	0.68	0.85	0.48	0.4	0.46
CFOREST1	0.87	0.1	0.82	0.28	1	1
CFOREST2	0.82	0.16	0.69	0.78	0.99	1
CTREE	0.4	0.07	0.65	0.54	0.99	1
GBM	0.8	0.14	0.62	0.88	0.99	0.98
GUIDE	0.99	0.3	0.94	0.24	0.9	0.92
LASSO	0.19	0.59	0.75	0.55	0.73	0.76
RANGER	0.97	0.18	0.96	0.33	1	1
RF	0.83	0.16	0.54	0.28	0.87	0.91
RFSRC	0.79	0.02	0.72	0.8	1	1
RLT	0.69	0	0.67	0.77	0.99	1
RPART	0.92	0.2	0.85	0.44	0.9	0.93

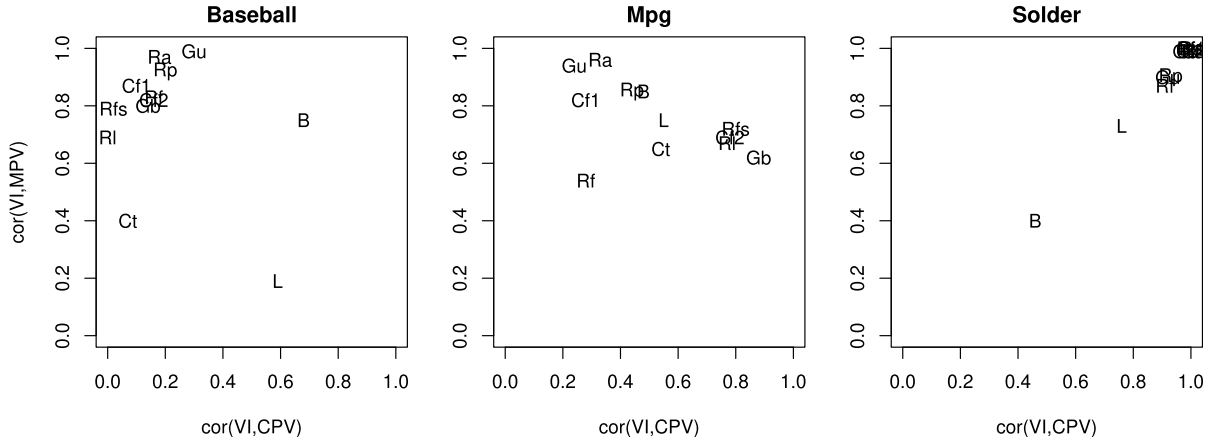

Figure 10: cor(`VI`, `MPV`) vs. cor(`VI`, `CPV`) for three data sets; `B` = BARTM, `Cf1` = CFOREST1, `Cf2` = CFOREST2, `Ct` = CTREE, `Gb` = GBM, `Gu` = GUIDE, `L` = LASSO, `Ra` = RANGER, `Rf` = RF, `Rfs` = RFSRC, `Rl` = RLT, `Rp` = RPART.

RFSRC (0.80) and CFOREST2 (0.78).

Solder. Owing to the almost perfect correlation between `MPV` and `CPV`, their correlations with the importance scores are essentially the same. BARTM and LASSO are the only two methods with correlations substantially below 0.90, suggesting that they measure something besides `MPV` and `CPV`.

Across the three data sets, the importance scores of all methods except for `BARTM` and `LASSO` are consistent with `MPV`, with GUIDE, RANGER and RPART showing the highest consistency. Consistency with `CPV` is weaker and more variable between data sets.

Table 5: Important variables (in alphabetical order for BARTM, in decreasing importance for GUIDE) for $\alpha = 0.05$.

Data	BARTM	GUIDE
COVID	diabetes, race=Black or African American, race=Unknown, race=White	renal, charlson, agecat, MI, CHF, dementia, PVD, cerebro, cancer, diabetes, race, CPD, sex, metastatic, hemipara, modsevliv, mildliver
Baseball	Hitcr, Rbcr, Runcr, Yrs	Batcr, Hitcr, Runcr, Rbcr, Wlkcr, Yrs, Hrcr, Hit86, Rb86, Bat86, Wlk86, Run86, Hr86, Pos86, Puto86
Mpg	Cylin=3, Cylin=4, Enginsz, Hp, Make=Honda, Make=Kia, Make=Toyota, Type=car, Weight	Weight, Enginsz, Cylin, Hp, Dcost, Rprice, Width, Whlbase, Drive, Type, Make, Length, Region
Solder	mask=B6, opening=small	opening, mask, solder, padtype

6 Thresholding

It is useful to have a score threshold to identify variables that are independent of the response. This is particularly desirable if the number of variables is large. Of the 12 scoring methods, only BARTM and GUIDE currently provide thresholds. We call a variable "unimportant" if it is independent of the response variable and "important" otherwise. Under the null hypothesis H_0 that all variables are unimportant, we define a "Type I error" as that of declaring at least one variable important. To control the probability of this error at significance level α, Bleich et al. (2014) randomly permute the Y values several times, keeping the X values fixed. They construct a BARTM forest to each set of permuted data, derive several candidate thresholds from the permutation distributions of the variable selection frequencies, and use cross-validation to choose among them.

GUIDE similarly permutes the Y values, keeping the X values fixed. For $j = 1, 2, \ldots, 300$, let u_j denote the maximum value of the GUIDE importance scores for the jth permuted data set and let $u^*(\alpha)$ be the $(1-\alpha)$-quantile of the set $\{u_1, u_2, \ldots, u_{300}\}$. Under H_0, the probability that at least one score exceeds the value of $u^*(\alpha)$ is approximately α.

Bias adjustment of the importance scores defined in equation (2) requires one level of permutation and calculation of $u^*(\alpha)$ requires another level. GUIDE uses an approximation to skip the second level. In the permutations for bias adjustment, let $v_b = \max_i v_b^*(X_i)$, $b = 1, 2, \ldots, B$, denote the maximum unadjusted score, where $v_b^*(X_i)$ is defined above equation (2). Let $v^*(\alpha)$ denote the $(1-\alpha)$-quantile of $\{v_1, v_2, \ldots, v_B\}$. Let $s(X_i)$ be the unadjusted score for the unpermuted (real) data defined in (1). Finally, let k denote the number of values of $s(X_i)$ greater than $v^*(\alpha)$. Declare as important the variables with the top k values of the bias-adjusted scores $\texttt{VI}(X_i)$. Let $\tilde{v}(\alpha)$ denote the average of the kth and $(k+1)$th largest values of $\texttt{VI}(X_i)$. GUIDE reports the normalized importance scores $\texttt{VI}(X_i)/\tilde{v}(\alpha)$, so that variables with normalized scores less than 1.0 are considered unimportant.

Table 5 lists the variables found to be important by BARTM and GUIDE in the COVID, Baseball, Mpg, and Solder data sets, using $\alpha = 0.05$. GUIDE orders the important variables by their VI values, but BARTM does not order them. The table shows that BARTM tends to find fewer important variables than GUIDE. Besides, because it transforms each categorical

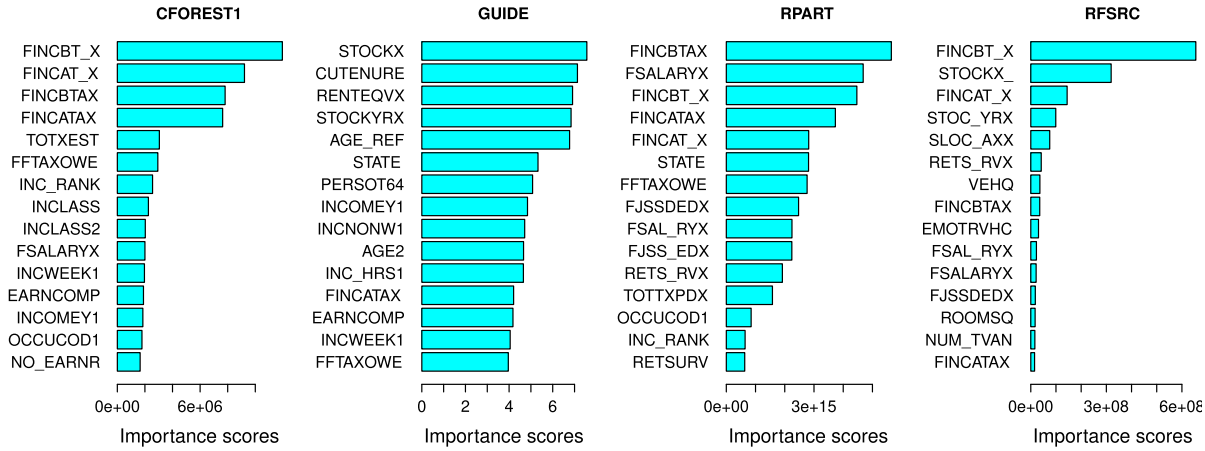

Figure 11: Variables with 15 highest importance scores in CE data.

variable into several indicator variables, BARTM may find some indicators important and other indicators unimportant. For example in the SOLDER data, BARTM finds only one of the four levels of `mask` and one of the three levels of `opening` to be important.

7 Missing Values

Among the 12 scoring methods, only CFOREST1, GUIDE, RPART, and RFSRC accept data with missing values. GUIDE importance scores are unbiased when there are missing values because the latter are treated as a special type of observation as described in Section 2. To show this and observe the effect of missing values on CFOREST1, RPART and RFSRC, we apply the methods to a data set from a Bureau of Labor Statistics 2013 Consumer Expenditure (CE) Survey that contains observations on more than 400 variables from 6464 respondents. We choose as dependent variable the amount of interest and dividends from the previous year (`INTRDVX`). About 25% of the values of `INTRDVX` are missing, either because the question is inapplicable or the respondent refused to answer it. For this demonstration, we use the 4693 respondents with non-missing `INTRDVX` to obtain importance scores for its prediction. Within this subset, about 20% of the other variables have missing values, with 67 of them having more than 95% missing, including `STOCKX` (value of directly-held stocks, bonds, mutual funds, etc.), which may be expected to be a good predictor of `INTRDVX`. See Loh et al. (2019, 2020) for more information on the variables.

Figure 11 shows barplots of the scores of the top 15 variables for each method. `STOCKX` is ranked most important by GUIDE and second most important by RFSRC, but it is not ranked in the top 15 by CFOREST1 and RPART. At least one of `FINCBTAX` (income before tax) or `FINCATAX` (income after tax) is in the top 15 of all four methods. These two variables have no missing values.

We can use the same procedure that produced Figure 2 to find out if there is bias in the importance scores by randomly permuting the `INTRDVX` values while holding the values of the predictor variables fixed. Let J be the number of permutations and $m_j(k)$ be the importance score of variable X_k in permutation j ($j = 1, 2, \ldots, J$). Figure 12 plots $\bar{m}(k) = J^{-1} \sum_j m_j(k)$ (arranged in increasing order) and their 2-SE error bars for each method, with $J = 1000$.

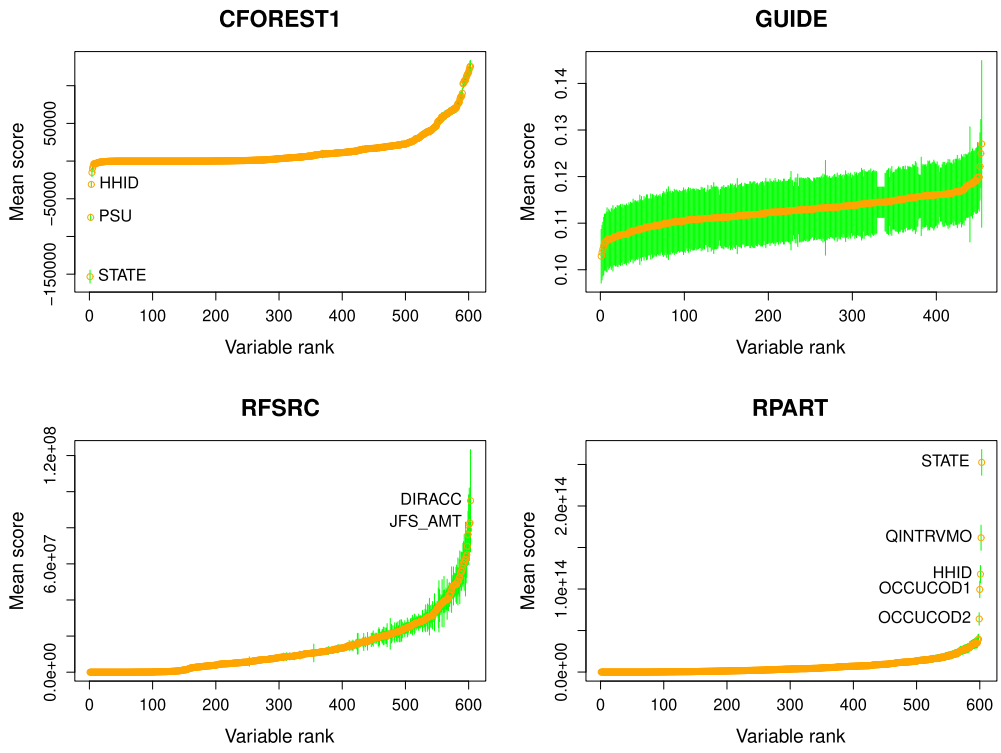

Figure 12: Mean importance scores $\overline{\texttt{VI}}$ and 2-SE bars from 1000 random permutations of the response variable for CE data. Variables ordered by increasing mean scores. GUIDE has fewer variables because it combines missing-value flag variables with their associated variables.

GUIDE is the only method with unbiased scores as evidenced by its overlapping 2-SE bars. The 2-SE bars in the other three methods seldom overlap (some bars are too short to be visible). CFOREST1 gives the lowest mean scores to categorical variables `HHID` (household identifier, 46 levels), `PSU`, (primary sampling unit, 21 levels), and `STATE` (39 levels)—due to its bias against non-dichotomous categorical variables (see Table 3). On the other hand, RPART gives very high mean scores to `STATE`, `HHID`, and categorical variables `OCCUCOD1` (respondent occupation, 15 levels) and `OCCUCOD2` (spouse occupation, 15 levels)—due to its bias towards non-dichotomous categorical variables. Finally, RFSRC gives its highest mean scores to binary variable `DIRACC` (access to living quarters) and continuous variable `JFS_AMT` (annual value of food stamps)—we have no explanation for this.

8 Conclusion

We have presented an improved importance scoring method based on the GUIDE algorithm and compared it with 11 methods in terms of selection bias and consistency with two measures of predictive importance. We say that a method is unbiased if the expected values of its scores are equal when all variables are independent of the response variable. We found that if the data do not have missing values, only CFOREST2, CTREE, GUIDE, and RANGER are unbiased. RF and RFSRC give lower scores to all categorical variables. GBM and RLT give higher scores to high-level categorical variables and lower scores to dichotomous variables. RPART gives lower

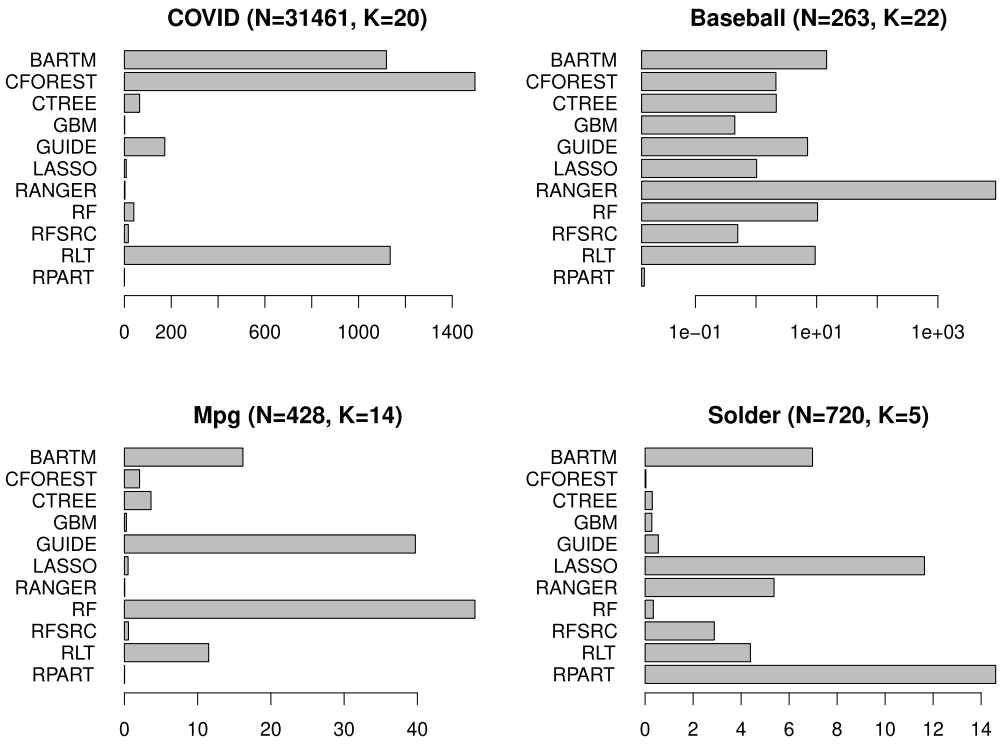

Figure 13: Average CPU times (sec.) for one set of importance scores; N is the sample size and K is the number of variables; plot for baseball data is on logarithmic scale.

scores to dichotomous variables. BARTM, CFOREST1 and LASSO have biases that are not easy to characterize. Only CFOREST1, GUIDE, RPART, and RFSRC are applicable to data with missing values, with GUIDE the only one that is unbiased. Unbiasedness of GUIDE is achieved through bias correction by random permutation of the values of the response variable. The technique is applicable to any scoring method that is not highly biased, but it can increase computational cost by an order of magnitude.

Figure 13 shows average computation times in seconds for each method to calculate one set of importance scores for the four data sets without missing values. Computations were performed on a 56-core Intel Xeon 2.40 GHz computer with 240 GB memory. The times are averages over 3 replications, to reduce the variability of randomized methods (CFOREST, GBM, LASSO, RANGER, RF, RFSRC, RLT) that employ random number seeds. In real applications the randomized methods will take much longer, because the number of replications need to be increased to stabilize the mean scores. The barplot for the Baseball data is drawn on a log scale due to the unusually long computation time for RANGER (we conjecture that this is due to the presence of 3 categorical variables each with 23 levels). Although the computation times for the Mpg and Solder data sets may not be large enough to be practically important, their relative sizes would be important if the sample sizes were much larger.

We used three data sets to examine whether the importance scores correlate well with two measures of predictive power, namely marginal predictive value (where other variables are ignored) and conditional predictive value (where other variables are fitted first). We found that the scores of many methods are highly correlated (> 0.80) with marginal predictive value, the exceptions being BARTM, CTREE, and LASSO. Correlations with conditional predictive values

Table 6: Score properties (✓indicates the method possesses the property). U: unbiased scores; C: 0.90 or higher correlation with marginal predictive values (MPV); T: threshold available; M: missing values allowed.

Method	U	C	T	M
BARTM			✓	
CFOREST1				✓
CFOREST2	✓			
CTREE	✓			
GBM				
GUIDE	✓	✓	✓	✓
LASSO				
RANGER	✓	✓		
RF				
RFSRC				✓
RLT				
RPART				✓

are generally low, except for CFOREST2, GBM, RFSRC, and RLT, where the correlations range from 0.77 to 0.88 in one data set.

Finally, we showed how GUIDE constructs $100(1-\alpha)\%$ threshold scores for distinguishing important from unimportant variables. The thresholds are constructed such that if all predictors are independent of the response, the probability that at least one score exceeds the thresholds is α. As with bias correction, the thresholding technique may be incorporated into other methods. Table 6 lists the properties of each method.

Supplementary Material

Data files and simulation programs used in the article may be found in a supplementary file.

Acknowledgement

The authors are grateful to the Editor, Associate Editor, and two referees for many helpful comments.

References

Bi J (2012). A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. *Journal of Sensory Studies*, 27: 87–101.

Bleich J, Kapelner A, George EI, Jensen ST (2014). Variable selection for BART: An application to gene regulation. *Annals of Applied Statistics*, 8: 1750–1781.

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

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton.

Bring J (1994). How to standardize regression coefficients. *American Statistician*, 48: 209–213.

Bureau A, Dupuis J, sK F, Lunetta KL, Hayward B, Keith TP, et al. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, 28: 171–182.

Chambers JM, Hastie TJ (1992). An appetizer. In: *Statistical Models in S* (JM Chambers, TJ Hastie, eds.), 1–12. Wadsworth & Brooks/Cole, Pacific Grove.

Chaudhuri P, Huang MC, Loh WY, Yao R (1994). Piecewise-polynomial regression trees. *Statistica Sinica*, 4: 143–167.

Chipman HA, George EI, McCulloch RE (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4: 266–298.

Denby L (1986). Major league baseball salary and performance data. http://lib.stat.cmu.edu/datasets/baseball.data.

Díaz-Uriarte R, Alvarez de Andrés S (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3): 3.

Friedman J (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29: 1189–1232.

Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22.

Greenwell B, Boehmke B, Cunningham J, Developers G (2019). *gbm: Generalized Boosted Regression Models*. R package version 2.1.5.

Harrison SL, Fazio-Eynullayeva E, Lane DA, Underhill P, Lip GYH (2020). Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United States: A federated electronic medical record analysis. *PLoS Medicine*, 17(9): 1–11.

Hoaglin DC, Velleman PF (1995). A critical look at some analyses of Major League Baseball salaries. *American Statistician*, 49: 277–285.

Hothorn T, Hornik K, Zeileis A (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15: 651–674.

Ishwaran H (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1: 519–537.

Ishwaran H, Kogalur U (2007). Random survival forests for R. *R News*, 7(2): 25–31.

Ishwaran H, Kogalur U, Blackstone E, Lauer M (2008). Random survival forests. *Annals of Applied Statistics*, 2(3): 841–860.

Johnson RW (2004). 2004 new car and truck data. http://jse.amstat.org/datasets/04cars.txt.

Kim H, Loh WY (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96: 589–604.

Kuhn M (2020). *caret: Classification and Regression Training*. R package version 6.0-86.

Liaw A, Wiener M (2002). Classification and regression by randomforest. *R News*, 2(3): 18–22.

Loh WY (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12: 361–386.

Loh WY (2009). Improving the precision of classification trees. *Annals of Applied Statistics*, 3: 1710–1737.

Loh WY (2012). Variable selection for classification and regression in large p, small n problems. In: *Probability Approximations and Beyond* (A Barbour, HP Chan, D Siegmund, eds.), volume 205 of *Lecture Notes in Statistics—Proceedings*, 133–157. Springer, New York.

Loh WY, Eltinge J, Cho MJ, Li Y (2019). Classification and regression trees and forests for incomplete data from sample surveys. *Statistica Sinica*, 29: 431–453.

Loh WY, Shih YS (1997). Split selection methods for classification trees. *Statistica Sinica*, 7: 815–840.

Loh WY, Vanichsetakul N (1988). Tree-structured classification via generalized discriminant analysis (with discussion). *Journal of the American Statistical Association*, 83: 715–728.

Loh WY, Zhang Q, Zhang W, Zhou P (2020). Missing data, imputation and regression trees. *Statistica Sinica*, 30: 1697–1722.

Lundberg SM, Lee SI (2017). A unified approach to interpreting model predictions. In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, eds.), 4768–4777.

Nembrini S, König IR, Wright MN (2018). The revival of the Gini importance? *Bioinformatics*, 21: 3711–3718.

Ribeiro MT, Singh S, Guestrin C (2016). "Why should I trust you?": Explaining the predictions of any classifier. In: *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Sandri M, Zuccolotto Z (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, 17: 611–628.

Strobl C, Boulesteix A, Kneib T, Augustin T, Zeileis A (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9: 307.

Strobl C, Boulesteix A, Zeileis A, Hothorn T (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8: 25.

Therneau TM, Atkinson EJ (2019a). An introduction to recursive partitioning using the RPART routines. R vignette. https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Therneau TM, Atkinson EJ (2019b). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.

Wei P, Lu Z, Song J (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering & Systems Safety*, 142: 399–432.

White AP, Liu WZ (1994). Bias in information-based measures in decision tree induction. *Machine Learning*, 15: 321–329.

Wright MN, Ziegler A (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1): 1–17.

Wu Y, Boos DD, Stefanski LA (2007). Variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*, 102: 235–243.

Zhu R (2018). *Reinforcement Learning Trees*. R package version 3.2.2.

Zhu R, Zeng D, Kosorok MR (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, 110: 1770–1784.