

# A General Model-Based Variable Importance Metric for Supervised Learning

Brandon M. Greenwell\*

Illumination Works

and

Andrew J. McCarthy

The Perduco Group

and

Bradley C. Boehmke

Department of Operational Sciences, Air Force Institute of Technology

October 2, 2017

## Abstract

In the era of "big data", it is becoming more of a challenge to not only build state-of-the-art predictive models, but also gain an understanding of what's really going on in the data. For example, it is often of interest to know which, if any, of the predictors in a fitted model are relatively influential on the predicted outcome. Fortunately, there are a few modern algorithms—like random forests and boosted decision trees—that have a natural way of ranking predictors in terms of some measure of importance or relative influence. Other algorithms—like neural networks and support vector machines—are not capable of doing so and model-free approaches are generally used to measure each predictor's importance. In this paper, we discuss a model-based approach to measuring predictor importance in any supervised learning setting.

*Keywords:* 3 to 6 keywords, that do not appear in the title

---

\*The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

# 1 Introduction

Complex supervised learning algorithms, like neural networks and support vector machines, are more common than ever in predictive analytics, especially when dealing with large observational databases that don't adhere to the strict assumptions imposed by traditional statistical techniques (e.g., multiple linear regression which typically assumes linearity, homoscedasticity, and normality). However, it can be challenging to understand the results of such complex models and explain them to management. Graphical displays such as variable importance plots (when available) and partial dependence plots (PDPs) offer a simple solution.

PDPs are low-dimensional graphical renderings of the prediction function  $\hat{f}(\mathbf{x})$  so that the relationship between the outcome and predictors of interest can be more easily understood. These plots are especially useful in explaining the output from "black box" models. While PDPs can be constructed for any predictor in a fitted model, variable importance scores are more difficult to define. What does it mean for a feature to be important? It's influence on the predicted outcome? Some methods—like random forests and other tree-based methods—have a natural way of defining variable importance. Unfortunately, this is not the case for other popular supervised learning algorithms like support vector machines. In this paper, we consider a partial dependence-based variable importance metric that can be used with any supervised learning algorithm.

## 2 Background

We are often confronted with the task of extracting knowledge from large databases. For this task we turn to various statistical learning algorithms which, when tuned correctly, can have state-of-the-art predictive performance. However, having a model that predicts well is only solving part of the problem. It is also desirable to extract information about the relationships uncovered by the learning algorithm. For instance, we often want to know which predictors, if any, are important by assigning some type of variable importance score to each feature. Once a set of influential features has been identified, the next step would be to summarize the functional relationship between each feature, or subset thereof, and

the outcome of interest. However, since most statistical learning algorithms are "black box" models, extracting this information is not always straightforward. Luckily, some learning algorithms have a natural way of defining variable importance.

## 2.1 Model-based approaches to variable importance

In a binary decision tree, at each node  $t$ , a single predictor is used to partition the data into two homogeneous groups. The chosen predictor is the one that maximizes some measure of improvement  $\hat{i}_t$ . The relative importance of predictor  $x$  is the sum of the squared improvements over all internal nodes of the tree for which  $x$  was chosen as the partitioning variable; see [Breiman et al. \(1984\)](#) for details. This idea also extends to ensembles of decision trees, such as boosting and random forest. In ensembles, the improvement score for each predictor is averaged across all the trees in the ensemble. Fortunately, due to the stabilizing effect of averaging, the improvement-based variable importance metric is often more reliable in large ensembles ([Hastie et al., 2009](#), pg. 368). Random forests offer an additional method for computing variable importance scores. The idea is to use the leftover out-of-bag (OOB) data to construct validation-set errors for each tree. Then, each predictor is randomly shuffled in the OOB data and the error is computed again. The idea is that if variable  $x$  is important, then the validation error will go up when  $x$  is perturbed in the OOB data. The difference in the two errors is recorded for the OOB data then averaged across all trees in the forest.

In multiple linear regression, the absolute value of the  $t$ -statistic is commonly used as a measure of variable importance. The same idea also extends to generalized linear models and nonlinear least squares. Multivariate adaptive regression splines (MARS), which were introduced in [Friedman \(1991\)](#), is an automatic and adaptive regression technique which can be seen as a generalization of multiple linear regression and generalized linear models. In the MARS algorithm, the contribution (or variable importance score) for each predictor is determined using a generalized cross-validation (GCV) statistic.

For neural networks, two popular methods for constructing variable importance scores are the Garson algorithm ([Garson, 1991](#)), later modified by [Goh \(1995\)](#), and the Olden algorithm ([Olden et al., 2004](#)). Both algorithms use the network's connection weights to

form the basis of the variable importance scores. The Garson algorithm determines variable importance by identifying all weighted connections between the nodes of interest. Olden’s algorithm, on the other hand, uses the product of the raw input-hidden and hidden-output connection weights between each input and output neuron and sums the product across all hidden neurons. This has been shown to outperform the Garson method in various simulations.

## 2.2 Filter-based approaches to variable importance

Filter-based approaches, which are described in [Kuhn and Johnson \(2013, chap. 18\)](#), do not make use of the fitted model to measure variable importance. They also do not take into account the other variables in the model.

For regression problems, a popular approach to measuring the variable importance of a numeric predictor  $x$  is to first fit a flexible nonparametric model between  $x$  and the target  $Y$ ; for example, the locally-weighted polynomial regression (LOWESS) method of [Cleveland \(1979\)](#). From this fit, a pseudo- $R^2$  measure can be obtained from the resulting residuals and used as a measure of variable importance. For categorical predictors, a different method based on standard statistical tests (e.g.,  $t$ -tests and ANOVAs) is employed; see [Kuhn and Johnson \(2013, chap. 18\)](#) for details.

For classification problems, an area under the ROC curve (AUC) statistic can be used to quantify predictor importance. The AUC statistic is computed by using the predictor  $x$  as input to the ROC curve. If  $x$  can reasonably separate the classes of  $Y$ , then that is a clear indicator that  $x$  is an important predictor (in terms of class separation) and this is captured in the corresponding AUC statistic. For problems with more than two classes, extensions of the ROC curve or a one-vs-all approach can be used.

## 2.3 Partial dependence plots

[Harrison and Rubinfeld \(1978\)](#) analyzed a data set containing suburban Boston housing data from the 1970 census that now permeates statistics literature. They sought a housing value equation using the assortment of variables provided; see [Harrison and Rubinfeld \(1978, Table IV\)](#) for a description of each variable. The assumptions usually assumed

for any regression-based analysis, such as normality, linearity, and constant variance, were clearly violated, but through an exhausting series of transformations, significance testing, and grid searches, they were able to build a model which fit the data reasonably well ( $R^2 = 0.81$ ). Their prediction equation is given in Equation (1).

$$\begin{aligned} \log(\widehat{MV}) = & 9.76 + 0.0063RM^2 + 8.98 \times 10^{-5}AGE - 0.19 \log(DIS) + 0.096 \log(RAD) \\ & - 4.20 \times 10^{-4}TAX - 0.031PTRATIO + 0.36(B - 0.63)^2 - 0.37 \log(LSTAT) \\ & - 0.012CRIM + 8.03 \times 10^{-5}ZN + 2.41 \times 10^{-4}INDUS + 0.088CHAS \\ & - 0.0064NOX^2. \end{aligned} \tag{1}$$

Using modern computing power, many supervised learning algorithms can fit large data sets in seconds, producing powerful, highly accurate models. The downfall of many of these machine learning algorithms, however, is decreased interpretability. Unlike prediction formulas such as Equation (1), many machine learning algorithms are left attempting to convey some relative measure of variable importance which provide insight, but cannot match the simplicity of Equation (1) and others like it.

Tree-based methods can naturally assign variable importance scores to each feature, but to gain even more insight, one can construct a partial dependence plot (PDP); see [Friedman \(2001\)](#) for details. Not only do PDPs visually convey the relationship between low cardinality subsets of the features (usually 1-3) and the response (while accounting for the average effect of the other predictors in the model), they can also be used to rank and scores the predictors in terms of their relative influence on the predicted outcome, as will be demonstrated in this paper. PDPs are particularly effective at explaining the output from "black box" models such as random forests and support vector machines.

Let  $\mathbf{x} = \{x_1, x_2, \dots, x_p\}$  represent the predictors in a model whose prediction function is  $\widehat{f}(\mathbf{x})$ . If we partition  $\mathbf{x}$  into an interest set,  $\mathbf{z}_s$ , and its complement,  $\mathbf{z}_c = \mathbf{x} \setminus \mathbf{z}_s$ , then the "partial dependence" of the response on  $\mathbf{z}_s$  is defined as

$$f_s(\mathbf{z}_s) = E_{\mathbf{z}_c} [\widehat{f}(\mathbf{z}_s, \mathbf{z}_c)] = \int \widehat{f}(\mathbf{z}_s, \mathbf{z}_c) p_c(\mathbf{z}_c) d\mathbf{z}_c, \tag{2}$$

where  $p_c(\mathbf{z}_c)$  is the marginal probability density of  $\mathbf{z}_c$ :  $p_c(\mathbf{z}_c) = \int p(\mathbf{x}) d\mathbf{z}_s$ . Equation (2)

can be estimated from a set of training data by

$$\bar{f}_s(\mathbf{z}_s) = \frac{1}{n} \sum_{i=1}^n \hat{f}(\mathbf{z}_s, \mathbf{z}_{i,c}), \quad (3)$$

where  $\mathbf{z}_{i,c}$  ( $i = 1, 2, \dots, n$ ) are the values of  $\mathbf{z}_c$  that occur in the training sample; that is, we average out the effects of all the other predictors in the model.

Constructing a PDP (3) in practice is rather straightforward. To simplify, let  $\mathbf{z}_s = x_1$  be the predictor variable of interest with unique values  $\{x_{11}, x_{12}, \dots, x_{1k}\}$ . The partial dependence of the response on  $x_1$  can be constructed as follows:

**for**  $i \in \{1, 2, \dots, k\}$  **do**

- (1) copy the training data and replace the original values of  $x_1$  with the constant  $x_{1i}$ ;
- (2) compute the vector of predicted values from the modified copy of the training data;
- (3) compute the average prediction to obtain  $\bar{f}_1(x_{1i})$ .

**end**

Plot the pairs  $\{x_{1i}, \bar{f}_1(x_{1i})\}$  for  $i = 1, 2, \dots, k$ .

**Algorithm 1:** A simple algorithm for constructing the partial dependence of the response on a single predictor  $x_1$ .

Algorithm (1) can be quite computationally intensive since it involves  $k$  passes over the training records, but fortunately, is easily parallelizable. In addition, Algorithm (1) can be extended to larger subsets of two or more features.

While PDPs are an invaluable tool in understanding the relationships uncovered by complex nonparametric models, they can be misleading in the presence of substantial interaction effects (Goldstein et al., 2015). To overcome this issue Goldstein et al. developed the concept of individual conditional expectation (ICE) plots. ICE plots display the estimated relationship between the response and a predictor of interest for each observation; In other words, skipping step 1 (c) in Algorithm (1). Consequently, the PDP for a predictor of interest can be obtained by averaging the corresponding ICE curves across all observations.

In Section 3.2, we show how to use partial dependence functions to evaluate the strength of potential interactions effects.

## 2.4 Boston housing data

For illustration, we will use a corrected version of the Boston housing data analyzed in Harrison and Rubinfeld (1978); the data are available from Statlib at [http://lib.stat.cmu.edu/datasets/boston\\_corrected.txt](http://lib.stat.cmu.edu/datasets/boston_corrected.txt). Using the R package `randomForest` (Liaw and Wiener, 2002), we fit a random forest with tuning parameter  $m_{try} = 6$  (chosen using 5-fold cross-validation) and 1000 trees. The model fit is reasonable, with an *out-of-bag* (pseudo)  $R^2$  of 0.89. The variable importance scores are displayed in Figure 1. Both plots indicate that the percentage of lower status of the population (`lstat`) and the average number of rooms per dwelling (`rm`) are highly associated with the median value of owner-occupied homes (`medv`). They also indicate that the proportion of residential land zoned for lots over 25,000 sq.ft (`zn`) has little association with `medv`.

The PDPs for these three variables are displayed in Figure 2. Notice how the PDP for `zn` is essentially flat, compared to the PDPs for `lstat` and `rm`. It is this notion of "flatness" which we will use as a basis to define our variable importance measure.

## 3 A model-based variable importance measure

The PDP for `zn` in Figure 2 is relatively flat, indicating that `zn` does not have much influence on the predicted median home value. In other words, the partial dependence values  $\bar{f}(x_i)$  ( $i = 1, 2, \dots, k$ ) display little variability. One might conclude that any variable for which the PDP is "flat" is likely to be less important than those predictors whose PDP varies across a wider range of the response.

Our notion of variable importance is based on any measure of the "flatness" of the partial dependence function. In general, we define

$$\text{imp}(x) = \text{flatness}(\bar{f}_s(z_s)),$$

where flatness is any measure of "flatness" of the curve. A simple and effective measure to use is the sample standard deviation for continuous predictors and the range statistic

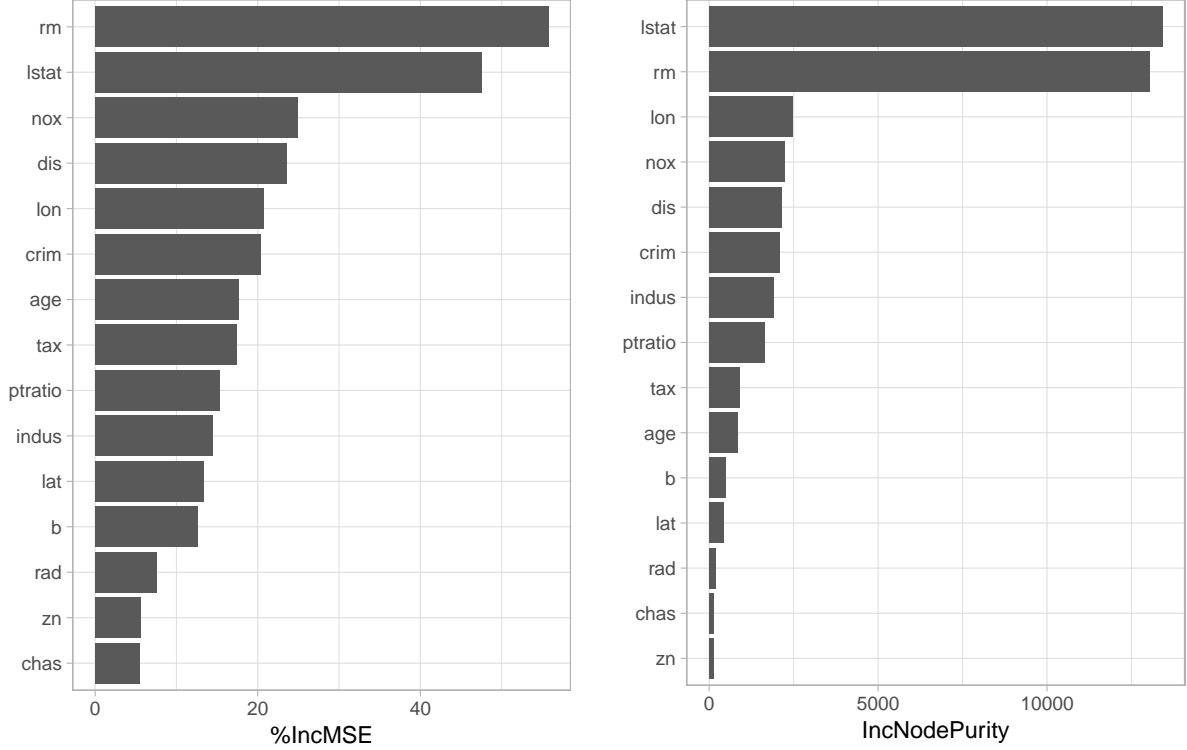


Figure 1: Variable importance scores from a random forest fit to the (corrected) Boston housing data. *Left*: OOB-based variable importance scores. *Right*: Impurity-based variable importance scores.

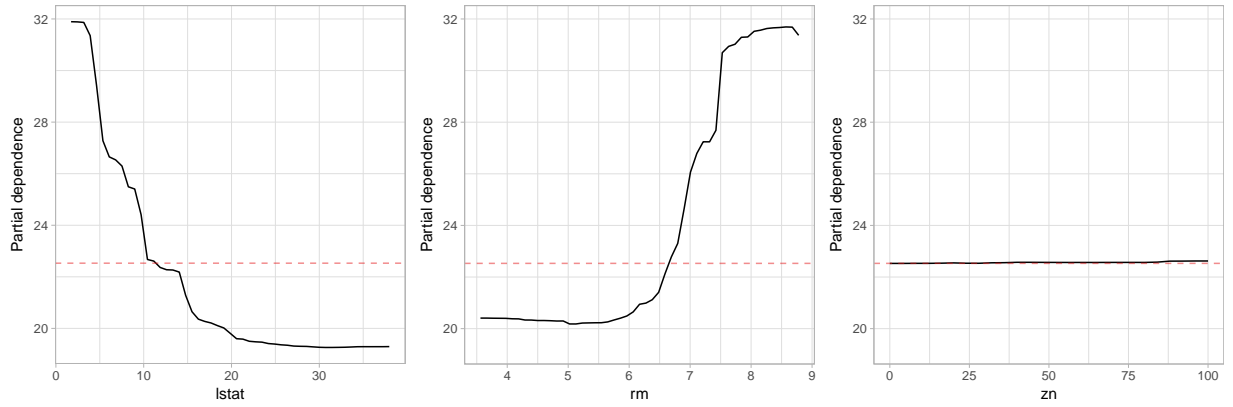


Figure 2: Partial dependence of  $\text{cmedv}$  on  $\text{lstat}$  (left),  $\text{rm}$  (middle), and  $\text{zn}$  (right). The mean of the  $n = 506$  median home values is indicated by a dashed red line.

(divided by 4) for factors with  $K$  levels. Based on Algorithm (1), our importance metric



for predictor  $x_1$  is simply

$$\text{imp}(x_1) = \sqrt{\frac{1}{k-1} \sum_{i=1}^k \left[ \bar{f}_1(x_{1i}) - \frac{1}{k} \sum_{i=1}^k \bar{f}_1(x_{1i}) \right]^2}, \quad (4)$$

if  $x_1$  is continuous and

$$\text{imp}(x_1) = \left[ \max_i (\bar{f}_1(x_{1i})) - \min_i (\bar{f}_1(x_{1i})) \right] / 4, \quad (5)$$

if  $x_1$  is categorical. Note that our variable importance metric relies on the fitted model; hence, it is crucial to properly tune and train the model to have good performance.

To illustrate, we apply Algorithm (1) to all of the predictors in the random forest for the Boston housing example and compute (4)<sup>1</sup>. The results are displayed in Figure 3. In this case, our partial dependence-based algorithm matches closely with the results from the random forest.

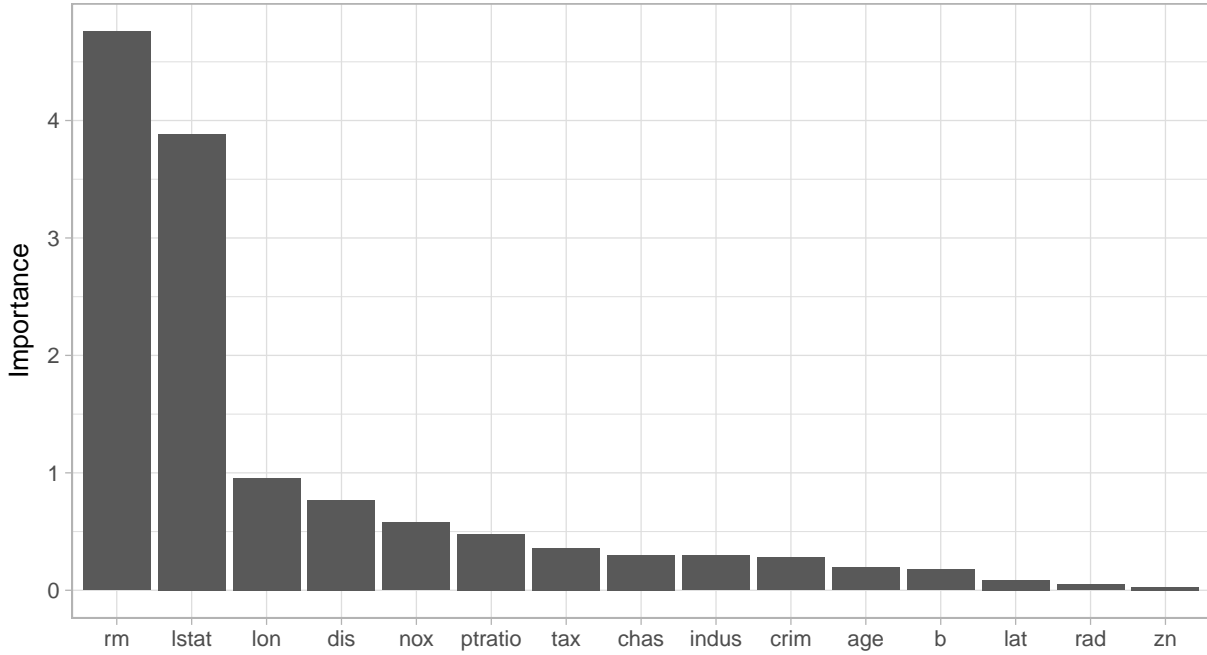


Figure 3: Partial-dependence-based variable importance scores from a random forest fit to the (corrected) Boston housing data.

<sup>1</sup>We used Equation (5) for `chas` since it is a categorical variable with two levels.

### 3.1 Linear models

As mentioned earlier, for linear models, a natural choice for measuring the importance of each term is to use the absolute value of the corresponding coefficient divided by its estimated standard error (i.e., the absolute value of the  $t$ -statistic). This turns out to be equivalent to the partial dependence-based metric (4) when the predictors are independently and uniformly distributed over the same range.

For example, suppose we have a linear model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

where  $\beta_i$  ( $i = 1, 2$ ) is a constant,  $X_1$  and  $X_2$  are both  $\mathcal{U}(0, 1)$  random variables, and  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ . Since we know the distribution of  $X_1$  and  $X_2$ , we can easily find  $f_1(X_1)$  and  $f_2(X_2)$ . For instance,

$$f_1(X_1) = \int_0^1 E[Y|X_1, X_2] p(X_2) dX_2,$$

where  $p(X_2) = 1$ . Simple calculus then leads to

$$f_1(X_1) = \beta_0 + \beta_2/2 + \beta_1 X_1 \quad \text{and} \quad f_2(X_2) = \beta_0 + \beta_1/2 + \beta_2 X_2.$$

Since  $E[Y|X_1, X_2] = f(X_1, X_2)$  is additive, the true partial dependence functions are just simple linear regressions in each predictor with their original coefficient and an adjusted intercept. Taking the variance of each gives

$$\text{Var}[f_1(X_1)] = \beta_1^2/12 \quad \text{and} \quad \text{Var}[f_2(X_2)] = \beta_2^2/12.$$

Hence, the standard deviations are just the absolute values of the original coefficients (scaled by the same constant).

To illustrate, we simulated  $n = 1000$  observations from the following linear model

$$Y = 1 + 3X_1 - 5X_2 + \epsilon,$$

where  $X_1$  and  $X_2$  are both  $\mathcal{U}(0, 1)$  random variables, and  $\epsilon \sim \mathcal{N}(0, 0.01^2)$ . For this example, we have

$$f_1(X_1) = -\frac{3}{2} + 3X_1 \quad \text{and} \quad f_2(X_2) = \frac{5}{2} - 5X_2.$$

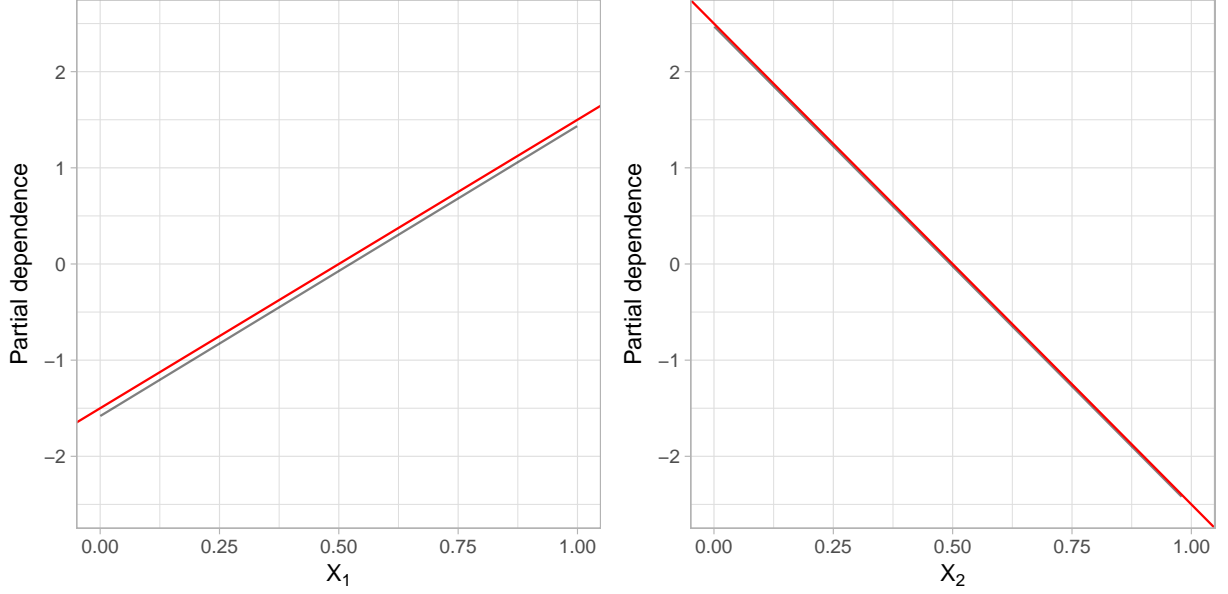


Figure 4: Estimated and true partial dependence plots.

These are plotted as red lines in Figure 4. Additionally, the black lines in Figure 4 correspond to the estimated partial dependence functions using Algorithm (1).

Clearly,  $X_2$  is more influential than  $X_1$ . Taking the absolute value of the ratio of the slopes in  $f_2(X_2)$  and  $f_1(X_1)$  gives  $5/3 \approx 1.67$ . In other words,  $X_2$  is roughly 1.67 times more influential on  $Y$  than  $X_1$ . Using the partial-dependence-based variable importance metric, we obtain  $\text{imp}(X_1) = 1.4828203$  and  $\text{imp}(X_2) = 0.8961719$  which gives the ratio  $\text{imp}(X_2)/\text{imp}(X_1) \approx 1.65$ . In fact, we can compute the true variance for both partial dependence functions:

$$\text{Var}[f_1(X_1)] = 3^2/12 \quad \text{and} \quad \text{Var}[f_2(X_2)] = 5^2/12.$$

Hence, the ratio of the true standard deviations is  $5/3 \approx 1.67$ .

Using the absolute value of the  $t$ -statistic becomes less useful in linear models when, for example, a predictor appears in multiple terms (e.g., interaction effects and polynomial terms). The partial dependence approach, on the other hand, does not suffer from such drawbacks.

## 3.2 Detecting interaction effects

The same idea can be applied to finding interaction effects. Let  $\text{imp}(x_i, x_j)$  ( $i \neq j$ ) be the standard deviation of the joint partial dependence values  $\bar{f}_{ij}(x_i, x_j)$  ( $i = 1, 2, \dots, k_i, j = 1, 2, \dots, k_j$ ). Essentially, a weak interaction effect of  $x_i$  and  $x_j$  on  $Y$  would suggest that  $\text{imp}(x_i, x_j)$  has little variation when either  $x_i$  or  $x_j$  is held constant while the other varies.

Let  $\mathbf{z}_s = (x_i, x_j)$ ,  $i \neq j$ , be any two predictors in the feature space  $\mathbf{x}$ . Construct the partial dependence function  $\bar{f}_s(x_i, x_j)$  and compute  $\text{imp}(x_i)$  for each unique value of  $x_j$ , denoted  $\text{imp}(x_i|x_j)$ , and take the standard deviation of the resulting importance scores. The same can be done for  $x_j$  and the results are averaged together. Large values (relative to each other) would be indicative of possible interaction effects.

## 3.3 Computational considerations

Constructing partial dependence functions using the brute force method described in Algorithm (1) can be computationally expensive. Some ideas are discussed in [Greenwell \(2017\)](#). Most importantly, Algorithm (1) is embarrassingly parallel and computing partial dependence functions for each predictor can be done rather quickly on a machine with a multi-core processor. For much larger data sets, it may be worthwhile to reduce the grid size by using specific quantiles for each predictor, rather than using all the unique observations. For example, the partial dependence function can be approximated very quickly by using the min and max of each predictor along with the deciles of the unique predictor values.

# 4 Friedman's regression problem

To further illustrate, we will use one of the regression problems described in Friedman (1991) and Breiman (1996). The feature space consists of ten independent  $\mathcal{U}(0, 1)$  random variables; however, only five out of these ten are actually used. The response is related to the features according to the formula

$$Y = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon,$$

where  $\epsilon \sim \mathcal{N}(0, \sigma)$ . Using the R package `nnet` (Venables and Ripley, 2002), we fit a neural network with one hidden layer containing eight units and a weight decay of 0.01 (these parameters were chosen using 5-fold cross-validation).

Variable importance plots are displayed in Figure 5. Notice how the Garson and Olden algorithms incorrectly label some of the features not in the true model as "important". For example, the Garson algorithm incorrectly labels  $x_8$  (which is not included in the true model) as more important than  $x_5$  (which is in the true model). Similarly, Olden's method incorrectly labels  $x_{10}$  as being more important than  $x_2$ . Our method, on the other hand, clearly labels all five of the predictors in the true model as being, by far, the most important in the fitted neural network.

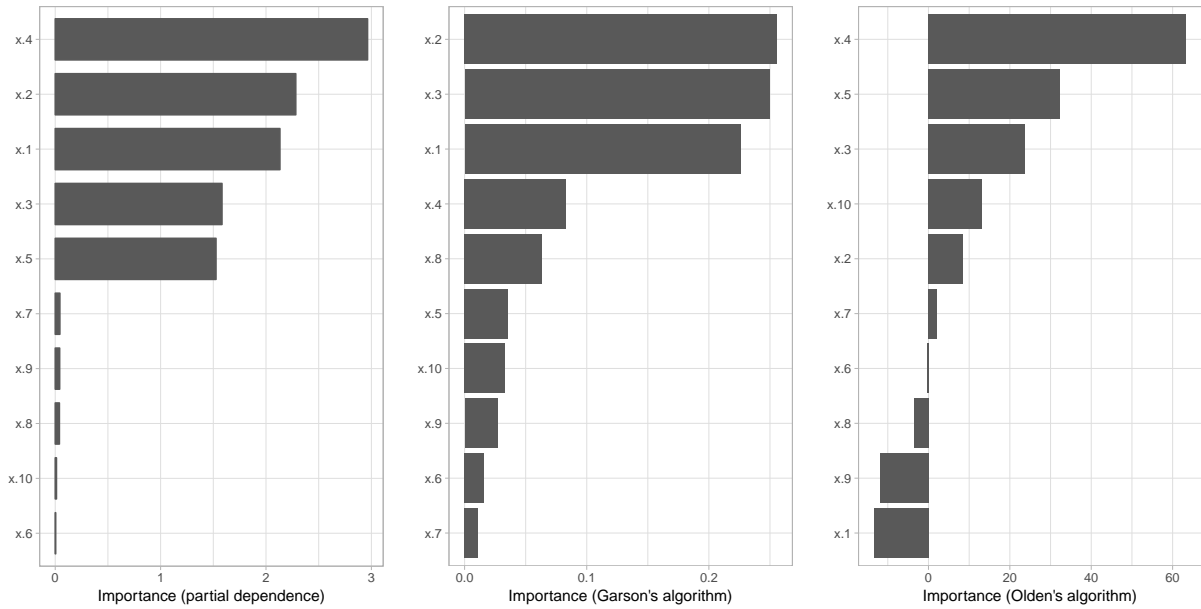


Figure 5: Variable importance plots neural network fit to the Friedman data set. *Left*: partial dependence-based algorithm. *Middle*: Garson's algorithm. *Right*: Olden's algorithm.

We also constructed the partial dependence functions for all pairwise interactions and computed the interaction statistic discussed in Section 3.2. The top ten interaction statistics are displayed in Figure 6. There is a clear indication of an interaction effect between the predictors  $x_1$  and  $x_2$ , the only interaction present in the true model.

In fact, since we know the distributions of the predictors in the true model, we can work

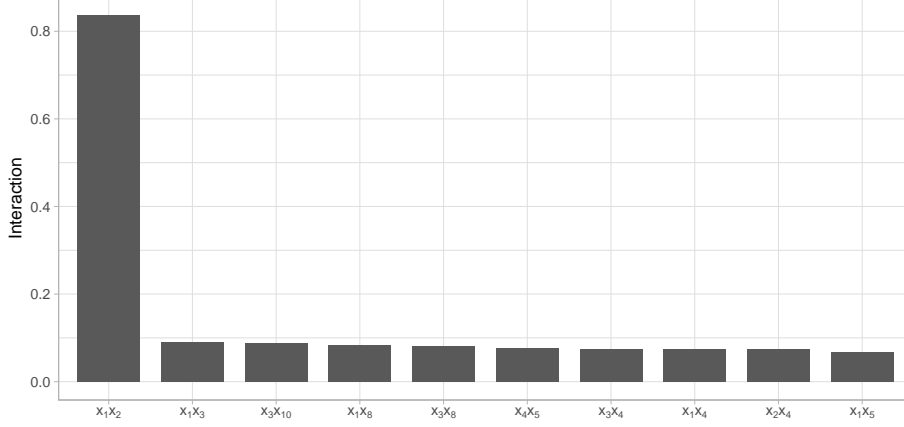


Figure 6: Variable importance-based interaction statistics from the neural network fit to the Friedman data set.

out the true partial dependence functions. For example, for the pairs  $(x_1, x_2)$  and  $(x_1, x_4)$ , we have

$$f(x_1, x_2) = 10 \sin(\pi x_1 x_2) + 55/6,$$

and

$$f(x_1, x_4) = \frac{5\pi x_1 (12x_4 + 5) - 12 \cos(\pi x_1) + 12}{6\pi x_1}.$$

Next, we simulated the standard deviation of  $f(x_1, x_2)$  for a wide range of fixed values of  $x_2$ ; this is what  $\text{imp}(x_1|x_2)$  is trying to estimate. Doing this for both predictors in each model resulted in Figure 7. The top row of Figure 7 illustrates that the importance of  $x_1$  (i.e., the strength of its relationship to the predicted outcome) heavily depends on the value of  $x_2$  and vice versa (i.e., an interaction effect between  $x_1$  and  $x_2$ ). In the bottom row, on the other hand, we see that the importance of  $x_1$  does not depend on the value of  $x_4$  and vice versa (i.e., no interaction effect between  $x_1$  and  $x_4$ ).

#### 4.0.1 Friedman's $H$ -statistic

An alternative measure for the strength of interaction effects is known as Friedman's  $H$ -statistic (Friedman and Popescu, 2008). Coincidentally, this statistic is also based on the partial dependence functions of the corresponding predictors, but uses a different approach.

For comparison, we fit a generalized boosted regression model (GBM) to the Fried-

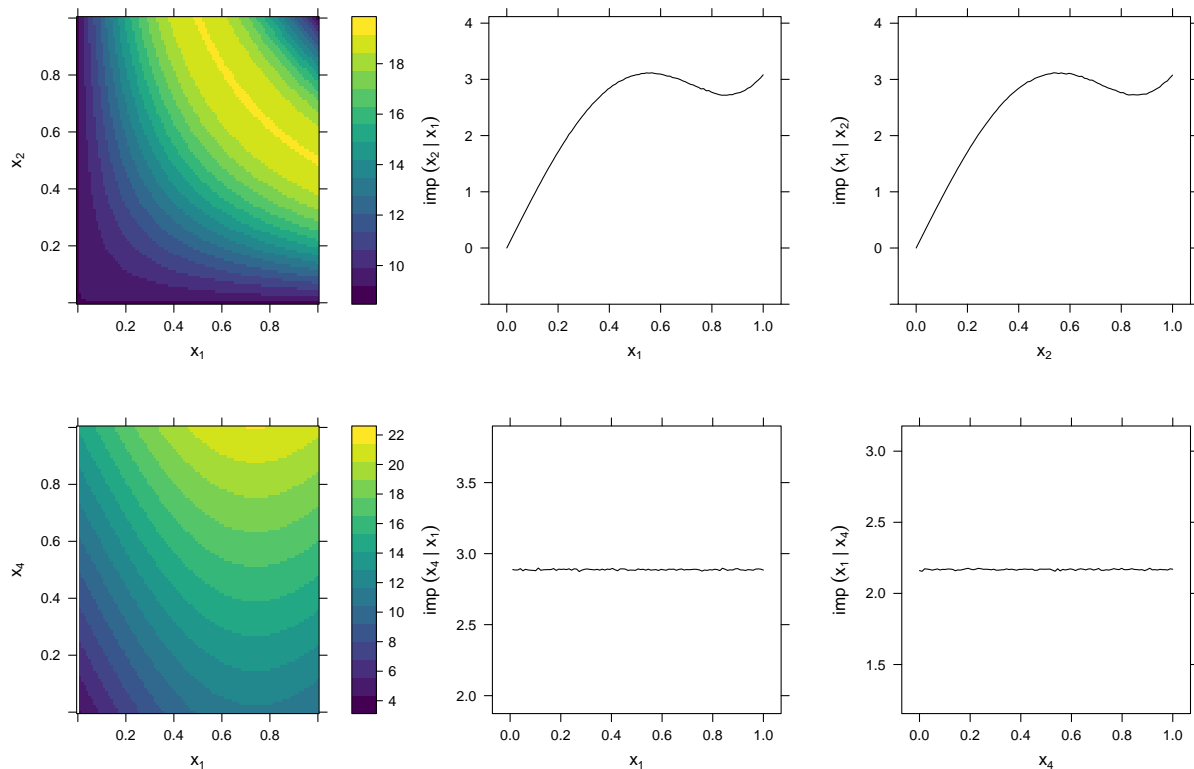


Figure 7: Results from a small Monte Carlo simulation on the interaction effects between  $x_1$  and  $x_2$  (top row), and  $x_1$  and  $x_4$  (bottom row).

man data set from the previous section. The parameters were chosen using 5-fold cross-validation. We used the R package `gbm` (Ridgeway, 2017) which has built-in support for computing Friedman’s  $H$ -statistic for any combination of predictors. The results are displayed in Figure 8. To our surprise, the  $H$ -statistic did not seem to catch the true interaction between  $x_1$  and  $x_2$ . Instead, the  $H$ -statistic ranked the pairs  $(x_8, x_9)$  and  $(x_7, x_{10})$  as having the strongest interaction effects, but these predictors do not even appear in the true model. Our variable importance-based interaction statistic, on the other hand, clearly suggests that the pair  $(x_1, x_2)$  has having the strongest interaction effect.

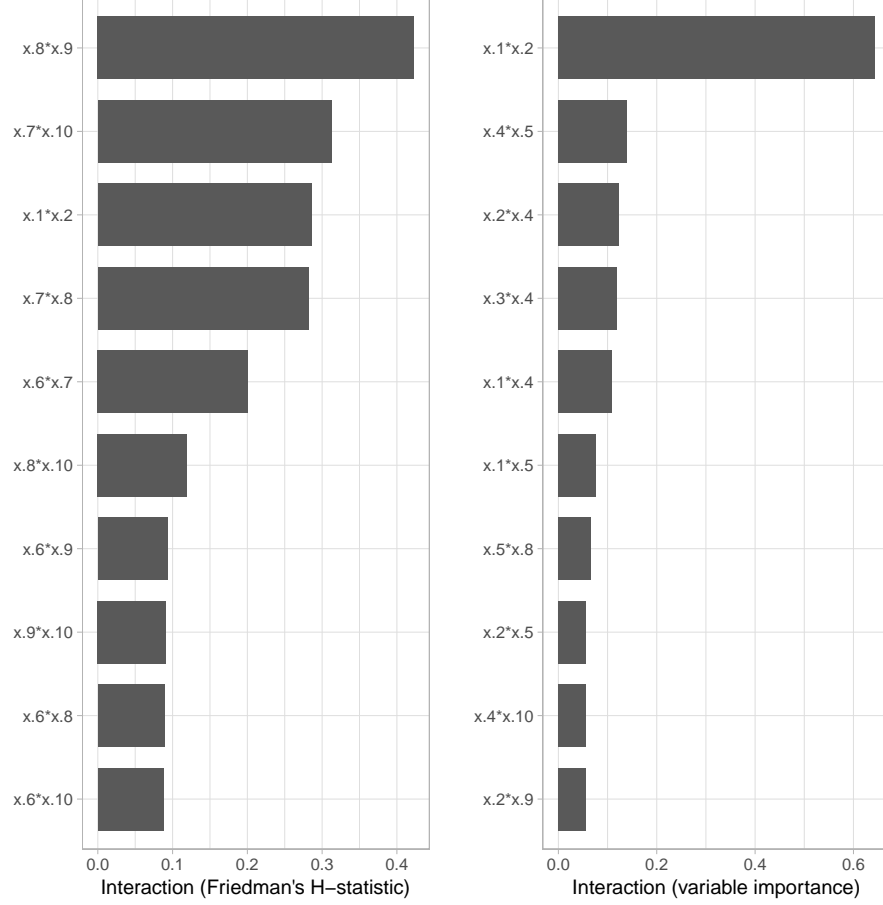


Figure 8: Interaction statistics for the GBM model fit to the Friedman data set. *Left*: Friedman’s  $H$ -statistic. *Right*: Our variable importance-based interaction statistic.

## 5 The Pima Indians diabetes data

In this section, we analyze the diabetes test results collected by the the US National Institute of Diabetes and Digestive and Kidney Diseases from a population of women who were at least 21 years old, of Pima Indian heritage, and living near Phoenix, Arizona. The data were taken directly from the `PimaIndiansDiabetes2` data frame which are available in the `mlbench` package (Leisch and Dimitriadou, 2012); for a description of the data and all the variables, see <https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes>.

We then used the `nnet` package (Venables and Ripley, 2002) in R to fit a neural network with seven hidden units and a weight decay of 0.01; these parameters were chosen to



maximize AUC using repeated 5-fold cross-validation with ten repeats; the cross-validated AUC was 0.77.

The partial dependence functions for all the predictors are displayed in Figure 9. The PDPs clearly indicate that plasma glucose concentration (**glucose**), diabetes pedigree function (**pedigree**), and 2-hour serum insulin (**insulin**) have the largest effect on the neural network’s prediction for the outcome of the diabetes test result (**diabetes**). On the other hand, the PDPs for diastolic blood pressure (**pressure**) and age (**age**) are relatively flat indicating they contribute less to the predictions than the other predictors. This is further illustrated in the variable importance plot displayed in Figure 10; the importance scores were obtained using Algorithm (4).

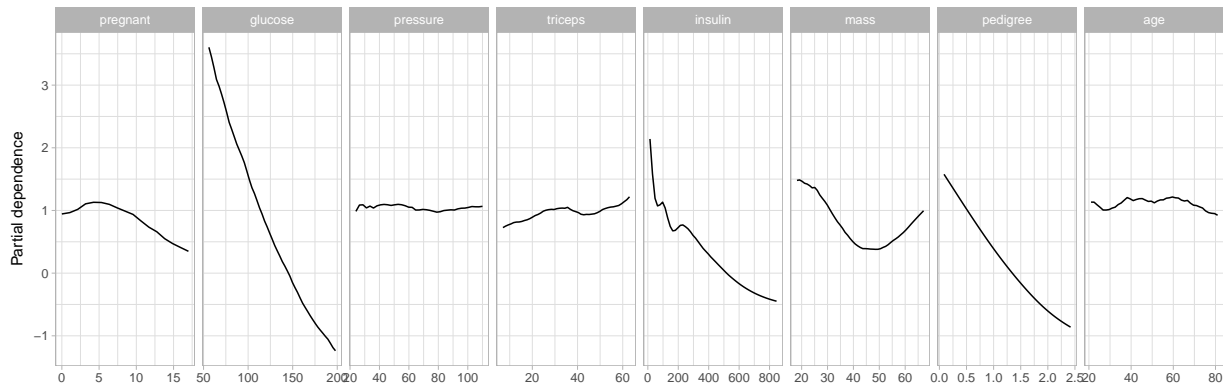


Figure 9: Partial-dependence plots for the eight predictors in the Pima Indians data set.

## 6 Conclusion

We have discussed a new partial dependence-based variable important metric for supervised learning. Since our algorithm is based on Friedman’s idea of partial dependence functions, our variable importance metric is model-based and takes into account the effects of the other features in the model. Consequently, this also requires that the fitted models be properly tuned to achieve optimum performance. We have also discussed a simple extension of our variable importance metric that seems useful for detecting interaction effects.

While this new approach to computing feature importance for any supervised learning algorithm seems very useful, more research is needed to determine where its deficiencies

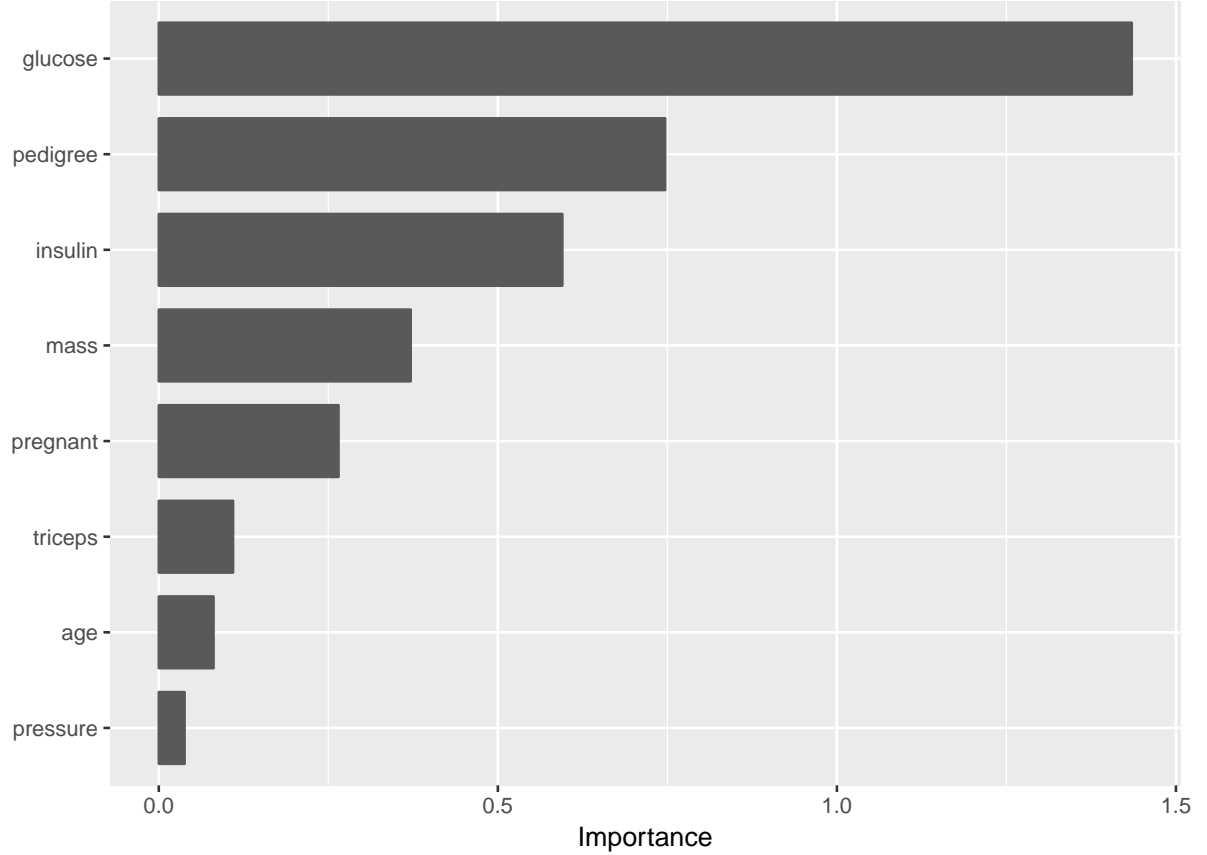


Figure 10: Partial-dependence-based variable importance scores for the eight predictors in the Pima Indians data set.

may lie. For example, outliers in the feature space can cause abnormally large fluctuations (outliers) in the partial dependence values  $\bar{f}(x_i)$  ( $i = 1, 2, \dots, k$ ). Therefore, it may be advantageous to use more robust measures of spread to describe the variability in the estimated partial dependence function; a reasonable choice would be the median absolute deviation which has a finite sample breakdown point of  $\lceil n/2 \rceil / n$ . It would also be interesting to see if our metric could be extended to rely on individual conditional expectation (ICE) curves ([Goldstein et al., 2015](#)), a refinement over Friedman’s PDP.

## SUPPLEMENTARY MATERIAL

**R-package vip:** Contains function for computing variable importance scores and constructing variable importance plots for various types of fitted models in R. The pack-

age is hosted on GitHub at <https://github.com/AFIT-R/vip>.

‘greenwell-vi-2017.R’ R script containing the R code to reproduce all of the results and figures in this paper.

## References

- Breiman, L., J. Friedman, and R. A. O. Charles J. Stone (1984). *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74(368), 829–836.
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* 19(1), 1–67.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 1189–1232.
- Friedman, J. H. and B. E. Popescu (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics* 2(3), 916–954.
- Garson, D. G. (1991). Interpreting neural-network connection weights. *Artificial Intelligence Expert* 6(4), 46–51.
- Goh, A. (1995). Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering* 9(3), 143–151.
- Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* 24(1), 44–65.
- Greenwell, B. M. (2017). pdp: An r package for constructing partial dependence plots. *The R Journal* 9(1), 421–436.

- Harrison, D. and D. L. Rubinfeld (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* 5(1), 81–102.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer-Verlag.
- Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling*. SpringerLink : Bücher. Springer New York.
- Leisch, F. and E. Dimitriadou (2012). *mlbench: Machine Learning Benchmark Problems*. R package version 2.1-1.
- Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News* 2(3), 18–22.
- Olden, J. D., M. K. Joy, and R. G. Death (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* 178(3), 389–397.
- Ridgeway, G. (2017). *gbm: Generalized Boosted Regression Models*. R package version 2.1.3.
- Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer-Verlag.