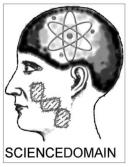
**British Journal of Mathematics & Computer Science**
**15(3): 1-9, 2016, Article no.BJMCS.23945**
*ISSN: 2231-0851*

**SCIENCEDOMAIN** *international*
www.sciencedomain.org




# Prediction Error Reduction Function as a Variable Importance Score

**E. Fokoué**[1*]

[1] *School of Mathematical Sciences, Rochester Institute of Technology, 98 Lomb Memorial Drive, Rochester, NY 14623, USA.*

***Author's contribution***

*The sole author designed, analyzed and interpreted and prepared the manuscript.*

***Article Information***

DOI: 10.9734/BJMCS/2016/23945
*Editor(s):*
(1) Carlo Bianca, Laboratoire de Physique Thorique de la Matire Condense, Sorbonne Universits, France.
*Reviewers:*
(1) Er. Anuj Goel, Maharishi Markandeshwar University, India.
(2) S. K. Srivatsa, Prathyusha Engg College, Chennai, India.
Complete Peer review History: http://sciencedomain.org/review-history/13781

***Original Research Article***

***Received: 29th December 2015***
***Accepted: 27th February 2016***
***Published: 21st March 2016***

## Abstract

This paper introduces and develops a novel variable importance score function in the context of ensemble learning, and demonstrates its appeal empirically. Our proposed score function is simple and more straightforward than its counterpart proposed in the context of random forest, and by avoiding permutations, it is by design computationally more efficient than the random forest variable importance function. Just like the random forest variable importance function, our score handles both regression and classification seamlessly. One of the distinct advantage of our proposed score is the fact that it offers a natural cut off at zero, with all the positive scores indicating importance and significance, while the negative scores are deemed indications of insignificance. An extra advantage of our proposed score lies in the fact it works very well beyond ensemble of trees and can seamlessly be used with any base learners in the random subspace learning context. Our examples, both simulated and real, demonstrate that our proposed score does compete mostly favorably with the random forest score.

**Corresponding author: E-mail: epfeqa@rit.edu;*

*Keywords: High-dimensional; variable importance; random subspace learning; out-of-bag error; random forest; large p small n; classification; regression; ensemble learning; base learner.*

**2010 Mathematics Subject Classification:** 62H30, 62H25.

# 1 Introduction

Consider a data set $\mathscr{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$ where $\mathbf{x}_i$ is a $p$-dimensional vector of attributes of potentially different types observable on some input space denoted here by $\mathscr{X}$, and $\mathbf{y}_i$ are the responses taken from $\mathscr{Y}$. We shall consider various scenarios, but mainly the regression scenario with $\mathscr{Y} = \mathbb{R}$ and the classification scenario with $\mathscr{Y} = \{1, 2, \cdots, K\}$. We consider the task of building the estimator $\widehat{f}(\cdot)$ of the true but unknown underlying $f$, and seek to build $\widehat{f}(\cdot)$ such that the true error (generalization error) is as small as possible. In this context, we shall use the average test error $\texttt{AVTE}(\cdot)$, as our measure of predictive performance, namely

$$\texttt{AVTE}(\widehat{f}) = \frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{m}\sum_{j=1}^{m}\ell(\mathbf{y}_j^{(r)}, \widehat{f}^{(r)}(\mathbf{x}_j^{(r)}))\right\}, \tag{1.1}$$

where $\left(\mathbf{x}_j^{(r)}, \mathbf{y}_j^{(r)}\right)$ is the $j$th observation from the test set at the $r$th random replication of the split of the data. Throughout this paper, we shall use the zero-one loss (1.2) for all our classification tasks.

$$\ell(\mathbf{y}_j^{(r)}, \widehat{f}^{(r)}(\mathbf{x}_j^{(r)})) = 1_{\{\mathbf{y}_j^{(r)} \neq \widehat{f}^{(r)}(\mathbf{x}_j^{(r)})\}} = \begin{cases} 1 & \text{if } \mathbf{y}_j^{(r)} \neq \widehat{f}^{(r)}(\mathbf{x}_j^{(r)}) \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

For regression tasks, we shall use the squared error loss (1.2), namely

$$\ell(\mathbf{y}_j^{(r)}, \widehat{f}^{(r)}(\mathbf{x}_j^{(r)})) = (\mathbf{y}_j^{(r)} - \widehat{f}^{(r)}(\mathbf{x}_j^{(r)}))^2. \tag{1.3}$$

Besides seeking the optimal predictive estimator of $f$, we also seek to select the most important (useful) predictor variables as a byproduct of our overall learning scheme. Indeed, while accurate prediction is very important in and of itself, it's often desirable or even crucial in some cases, provide the added description of the importance of the variables involved in the prediction task. The statistical literature is filled with thousands of papers on variable selection and measurement of variable importance. [1] and [2] propose a measure of variable importance for random forest. [3] and [4] have written many interesting contributions to the estimation of variable importance specifically in the context of classification and regression trees and their random forest ensemble learning extension. [5] implements some of those measures of variable importance for tree-based learning in their wonderful R package caret. [6] give a nice summary of variable importance and [7] also touches on this subject. [8] proposed a bias-correction improvement to the above measures of variable importance. Unlike all these authors whose work on variable importance measures concentrated solely on tree-based models, we herein propose a measure that goes beyond tree-based methods.

# 2 Construction of the Prediction Error Reduction Function

## 2.1 Definitions and tools for defining the score function

We consider the common framework of a $p$-dimensional input space $\mathscr{X}$ with typical input vector $\mathbf{x} = (\mathtt{x}_1, \cdots, \mathtt{x}_p)^\top$. We also consider building different models with different subsets of the $p$ original

variables. Let $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)^\top$ denote the $p$-dimensional indicator such that

$$\gamma_j = \begin{cases} 1 & \text{if } \mathtt{x}_j \text{ is active in the current model indexed by } \boldsymbol{\gamma} \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Assume that we are given an ensemble (collection or aggregation) of models, say

$$\mathscr{H} = \{\mathtt{g}(\cdot, \boldsymbol{\gamma}^{(1)})), \mathtt{g}(\cdot, \boldsymbol{\gamma}^{(2)})), \cdots, \mathtt{g}(\cdot, \boldsymbol{\gamma}^{(B)}))\} \tag{2.2}$$

where $\mathtt{g}(\cdot, \boldsymbol{\gamma}^{(b)}))$ denotes the function built with only those variables that are active in the $b$th model of the ensemble (aggregation), and $\boldsymbol{\gamma}^{(b)} = (\gamma_1^{(b)}, \cdots, \gamma_p^{(b)})$ with

$$\gamma_j^{(b)} = \begin{cases} 1 & \text{if } \mathtt{x}_j \text{ is active in the } b\text{-th model of the ensemble} \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

For instance, we may consider a homogeneous ensemble, i.e, an ensemble in which all the functions are of the same family, like the case where all the base learners are multiple linear regression (MLR) models differing by the variables upon which they are built. Consider a score function $\mathtt{score}(\mathtt{g}(\cdot, \boldsymbol{\gamma}^{(b)}))$ used to assess the performance of model indexed by the variables active in $\boldsymbol{\gamma}^{(b)}$. We propose a variable importance score in the form of a function that measures the importance of a variable $\mathtt{x}_j$ in terms of the reduction in average score

$$\mathtt{PERF}(\mathtt{x}_j) = \frac{1}{B}\sum_{b=1}^{B} \mathtt{score}(\mathtt{g}(\cdot, \boldsymbol{\gamma}^{(b)})) - \frac{1}{B_j}\sum_{b=1}^{B} \gamma_j^{(b)} \mathtt{score}(\mathtt{g}(\cdot, \boldsymbol{\gamma}^{(b)})) \tag{2.4}$$

where $B_j$ is the number of models containing the variable $\mathtt{x}_j$, specifically $B_j = \sum_{b=1}^{B} 1_{\{\gamma_j^{(b)}=1\}}$. In words,

$$\mathtt{PERF}(\mathtt{x}_j) = \mathtt{Average\ score\ over\ all\ models} - \mathtt{Average\ score\ over\ all\ models\ with\ x}_j$$

## 2.2 Properties and benefits of the PERF score function

Intuitively, $\mathtt{PERF}(\mathtt{x}_j)$ somewhat measures the impact of variable $\mathtt{x}_j$. In a way similar to the approach used by sports writers to decide the Most Valuable Player (MVP) on a team or in a league, $\mathtt{PERF}(\mathtt{x}_j)$ looks at the overall performance of the whole ensemble and then for each variable $\mathtt{x}_j$ computes the direction and magnitude of the change to that overall performance of the ensemble brought by its presence in models. In a sense, the variable with the highest PERF score is the Most Valuable Predictor (MVP) variable. In other words, if

$$j^\star = \underset{j=1,\cdots,p}{\mathtt{argmax}} \{\mathtt{PERF}(\mathtt{x_j})\}, \quad \text{then} \quad \mathtt{x}_{j^\star} = \mathtt{MVP}(\mathtt{x}_1, \cdots, \mathtt{x}_j, \cdots, \mathtt{x}_p).$$

From the above definition, it follows that *If a variable* $\mathtt{x}_j$ *is important, then its presence in any model will cause that model to perform better in the sense of having a lower than common average error (score). The average score of all models containing an important variable* $\mathtt{x}_j$ *should therefore be lower than the overall average score.* Essentially, (i) $|\mathtt{PERF}(\mathtt{x}_j)|$ measures the magnitude of the importance/impact. (ii) $\mathtt{sign}(\mathtt{PERF}(\mathtt{x}_j))$ measures the direction of the impact. (iii) If $\mathtt{sign}(\mathtt{PERF}(\mathtt{x}_j)) = +1$ and $|\mathtt{PERF}(\mathtt{x}_j)|$ is relatively large, then $\mathtt{x}_j$ is an important variable. Some of the benefits of the PERF score include the following: (a) The PERF score is seamlessly applied to both large $p$ small $n$ and small $p$ large $n$ machine learning settings, whether it be a classification or a regression task. (b) All variables with $\mathtt{PERF}(\mathtt{x}_j) \leq 0$ are unimportant and can be discarded. (c) The $\mathtt{PERF}(\cdot)$ score can be used whenever an ensemble $\mathscr{H}$ is available along with a suitable score function for each base learner. (d) This works with any base learner and can be adapted to parametric, nonparametric and semi-parametric models and one can imagine ensembles with any base learners as its atoms. (e) A great advantage over the traditional variable importance [1], [2] score functions is that the clear cut-off at zero, in the sense that all variables with $\mathtt{PERF}(\mathtt{x}_j) > 0$ are kept and all those variables with $\mathtt{PERF}(\mathtt{x}_j) \leq 0$ are thrown away.

## 2.3 Random subspace learning estimation of PERF

A natural implementation of $\texttt{PERF}(\cdot)$ can be done using the ubiquitous bootstrap along with the random subspace learning scheme. The `Out-of-Bag (oob)` error in the bagging or random subspace learning context is a good (in fact excellent) candidate score function, especially when the goal if the selection of variables that lead to the lowest prediction error. The advantage of using `oob` as the score lies in the fact that the score is obtained as part of building the ensemble in the random subspace learning framework. Consider the training set $\mathscr{D} = \{\mathbf{z}_i = (\mathbf{x}_i^\top, \mathrm{y}_i)^\top,\ i = 1, \cdots, n\}$, where $\mathbf{x}_i^\top = (\mathtt{x}_{i1}, \cdots, \mathtt{x}_{ip})$ and $\mathrm{y}_i \in \mathscr{Y}$ are realizations of two random variables $X$ and $Y$ respectively. Let $\mathbf{x}_{i,\boldsymbol{\pi}_j} = (\mathtt{x}_{i,1}, \cdots, \mathtt{x}_{i,\boldsymbol{\pi}_j}, \cdots, \mathtt{x}_{i,d})$. The permutation $\boldsymbol{\pi}_j$ acts the $|\bar{\mathscr{D}}^{(b)}|$-dimensional $j$th column of the out-of-bag data matrix. Essentially, $\boldsymbol{\pi}_j$ simply permutes the $|\bar{\mathscr{D}}^{(b)}|$ elements of the $j$th column of the out-of-bag data matrix.

**Algorithm 2.1.**

`Choose a base learner` $\widehat{\mathtt{g}}(\cdot)$ ▷ *e.g.: Trees, MLR*

`Choose an estimation method` ▷ *e.g.: Recursive Partitioning or OLS*

`Initialize all the` $\widehat{\texttt{PERF}}(\mathtt{x}_j)$ `and` $\widehat{\texttt{VI}}(\mathtt{x}_j)$ `at zero`

***for*** $b = 1$ *to* $B$ ***do***

`Draw with replacement` *from* $\mathscr{D}$ *a bootstrap sample* $\mathscr{D}^{(b)} = \{\mathbf{z}_1^{(b)}, \cdots, \mathbf{z}_n^{(b)}\}$

`Draw without replacement` *from* $\{1, \cdots, p\}$ *a subset* $\mathscr{V}^{(b)} = \{j_1^{(b)}, \cdots, j_d^{(b)}\}$ *of* $d$ *variables.*

`Form the indicator vector` $\boldsymbol{\gamma}^{(b)} = (\gamma_j^{(b)}, \cdots, \gamma_p^{(b)})$ *with*

$$\gamma_j^{(b)} = \begin{cases} 1 & \text{if } j \in \{j_1^{(b)}, \cdots, j_d^{(b)}\} \\ 0 & \text{otherwise.} \end{cases}$$

`Drop unselected variables` *from* $\mathscr{D}^{(b)}$ *so that* $\mathscr{D}_{\texttt{sub}}^{(b)}$ *is* $d$ *dimensional*

`Build the` $b$`th base learner` $\widehat{\mathtt{g}}(\cdot, \boldsymbol{\gamma}^{(b)})$ *based on* $\mathscr{D}_{\texttt{sub}}^{(b)}$

`Compute score of the` $b$`th base learner` $\widehat{\mathtt{g}}(\cdot, \boldsymbol{\gamma}^{(b)})$ ▷ *e.g. Out-of-bag error*

$$\mathtt{s}^{(b)} = \texttt{score}(\widehat{\mathtt{g}}(\cdot, \boldsymbol{\gamma}^{(b)})) = \frac{1}{|\bar{\mathscr{D}}^{(b)}|} \sum_{\mathbf{z}_i \notin \mathscr{D}^{(b)}} \ell(\mathrm{y}_i, \widehat{\mathtt{g}}(\mathbf{x}_i, \boldsymbol{\gamma}^{(b)}))$$

***for*** $j \in \mathscr{V}^{(b)}$ ***do***

`Generate the permutation of the` $j$`th column of` $\bar{\mathscr{D}}^{(b)}$`, namely`

$$\boldsymbol{\pi}_j$$

`Compute the permutation impacted score`

$$\mathtt{s}_{\boldsymbol{\pi}_j}^{(b)} = \texttt{score}_{\boldsymbol{\pi}_j}(\widehat{\mathtt{g}}(\cdot, \boldsymbol{\gamma}^{(b)})) = \frac{1}{|\bar{\mathscr{D}}^{(b)}|} \sum_{\mathbf{z}_i \notin \mathscr{D}^{(b)}} \ell(\mathrm{y}_i, \widehat{\mathtt{g}}(\mathbf{x}_{i,\boldsymbol{\pi}_j}, \boldsymbol{\gamma}^{(b)}))$$

`Compute the` $b$`th instance of the importance of` $\mathtt{x}_j$

$$\widehat{\texttt{VI}}^{(b)}(\mathtt{x}_j) = \mathtt{s}^{(b)} - \mathtt{s}_{\boldsymbol{\pi}_j}^{(b)}$$

***end for***

***end for***

*Use the ensemble* $\mathscr{H} = \left\{\widehat{\mathtt{g}}(\cdot, \boldsymbol{\gamma}^{(b)}),\ b = 1, \cdots, B\right\}$ *to form the estimator*

$$\widehat{\texttt{PERF}}(\mathtt{x}_j) = \frac{1}{B} \sum_{b=1}^{B} \texttt{score}(\widehat{\mathtt{g}}(\cdot, \boldsymbol{\gamma}^{(b)})) - \frac{1}{B_j} \sum_{b=1}^{B} \gamma_j^{(b)} \texttt{score}(\widehat{\mathtt{g}}(\cdot, \boldsymbol{\gamma}^{(b)})) \tag{2.5}$$

$$\widehat{\texttt{VI}}(\mathtt{x}_j) = \frac{1}{B_j} \sum_{b=1}^{B} \gamma_j^{(b)} \widehat{\texttt{VI}}^{(b)}(\mathtt{x}_j) \tag{2.6}$$

# 3 Computational Demonstrations

We herein assess the goodness and usefulness of the PERF variable importance score by applying to both simulated data and real life data. We specifically use benchmark machine learning data sets like the the spam detection dataset and the pima indian diabetes dataset in the classification context and the attitude dataset in the regression context. Our first example features simulated data with different scenarios on the level of correlation among the variables, and the ratio $n$ and $p$. In this particular example, the true function is

$$f(\mathbf{x}) = 1 + 2\mathtt{x}_3 + \mathtt{x}_7 + 3\mathtt{x}_9$$

with $\mathbf{x} \sim \mathtt{MVN}(\mathbf{1}_9, \Sigma_\rho)$ and $\epsilon \sim \mathbf{N}(0, 1)$. The dataset in this example is simulated data with different scenarios on the level of correlation among the variables, and the ratio $n$ and $p$. Specifically, we simulate data by defining $\rho \in [0, 1)$, then we generate our predictor variables using a multivariate normal distribution. Throughout this paper, the multivariate Gaussian density will be denoted by $\phi_p(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$

$$\phi_p(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \tag{3.1}$$

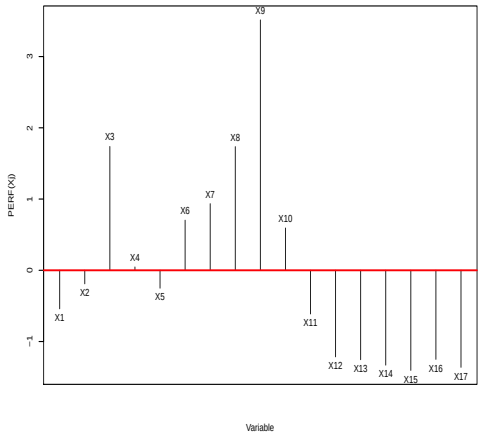
Furthermore, in order to study the effect of the correlation pattern, we simulate the data using a covariance matrix $\Sigma$ parameterized by $\tau$ and $\rho$ and defined by $\tau\Sigma$ where $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$.

$$\Sigma = \Sigma(\tau, \rho) = \tau \begin{pmatrix} 1 & \rho & \cdots & \rho^{p-2} & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{p-2} & \ddots & \rho & 1 & \rho \\ \rho^{p-1} & \rho^{p-2} & \cdots & \rho & 1 \end{pmatrix}$$

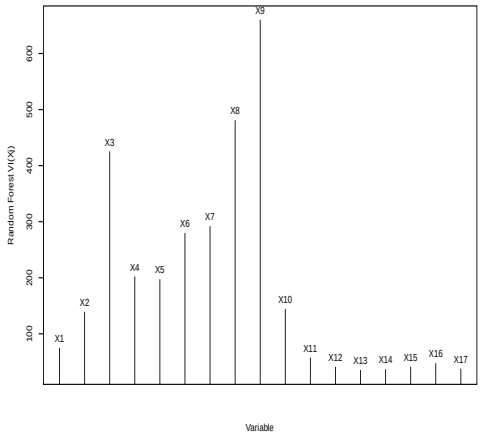
For simplicity however, we use the first $\Sigma$ with $\tau = 1$ throughout this paper. For the remaining parameters, we use $\rho \in \{0, 0.25, 0.75\}$ and $p \in \{17, 250\}$, with the same $n = 200$. The plots depicting the comparisons between Random Forest and PERF variable importance scores are given in subsequent parts of this paper. In each plot, the dimensionalities of the data are appropriately indicated, and comments are provided as well.

# 4 Discussion and Conclusion

We have presented a variable importance score function in the context of ensemble learning. Our proposed score function is simple and more straightforward than its counterpart proposed in the context of random forest, and by avoiding permutations, it is by design computationally more efficient than the random forest variable importance function. Just like the random forest variable importance function, our score handles both regression and classification seamlessly. One of the distinct advantage of our proposed score is the fact that it offers a natural cut off at zero, with all the positive scores indicating importance and significance, while the negative scores are deemed indications of insignificance. An extra advantage of our proposed score lies in the fact it works very well beyond ensemble of trees and can seamlessly be used with any base learners in the random subspace learning context. Our examples, both simulated and real, demonstrated that our proposed score does compete mostly favorably with the random forest score. In our future work, we present and compare the corresponding average test errors of the single models made up of the most important variables. We also provide in our future work theoretical proofs of the connection between our score function and the significance of variables selected using existing criteria. It is also our plan to address the fact that sometimes the correlation structure among the predictor variables obscures the ability of our proposed score to correctly identify some significant variables.

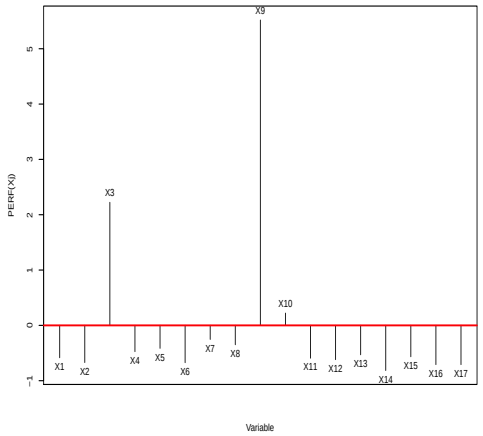



(a) Permutation-free Variable Importance.

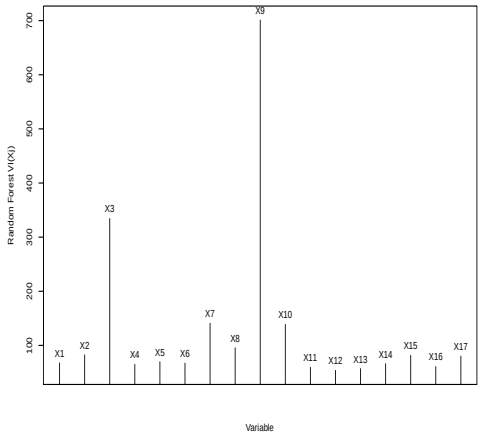



(b) Permutation-based Variable Importance.

**Fig. 1. Variable score for simulated data with high correlation among the variables in low dimension high sample size setting**

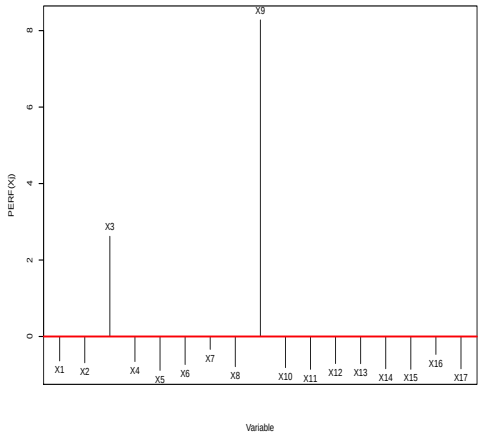



(c) Permutation-free Variable Importance.

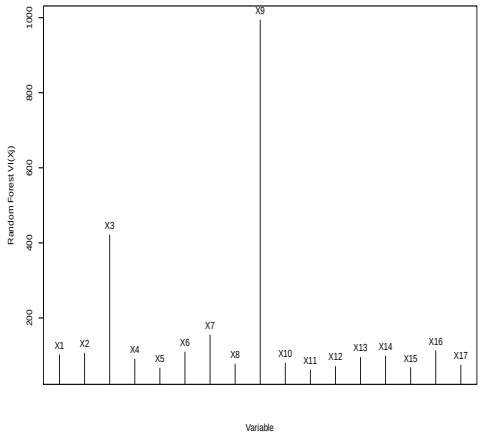



(d) Permutation-based Variable Importance.

**Fig. 2. Variable Importance Scores for simulated data with mild correlation among the variables in low dimension high sample size setting**

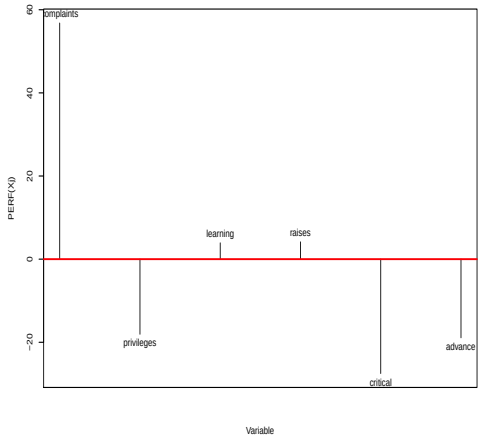



(e) Permutation-free Variable Importance.

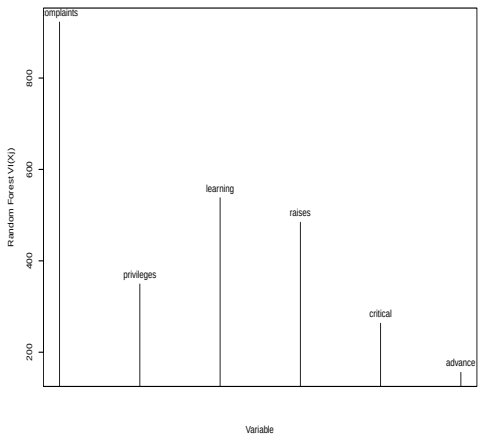



(f) Permutation-based Variable Importance.

**Fig. 3. Variable Importance Scores for simulated data with zero correlation among the variables in low dimension high sample size setting**

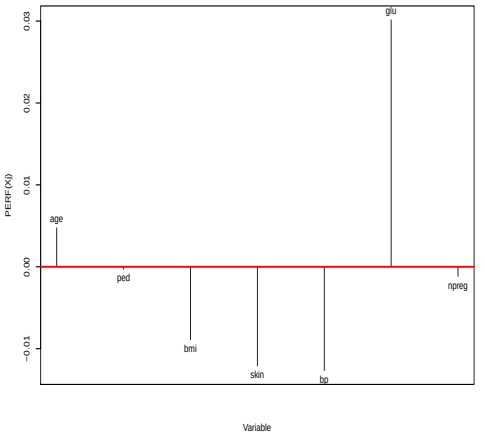



(g) Permutation-free Variable Importance.

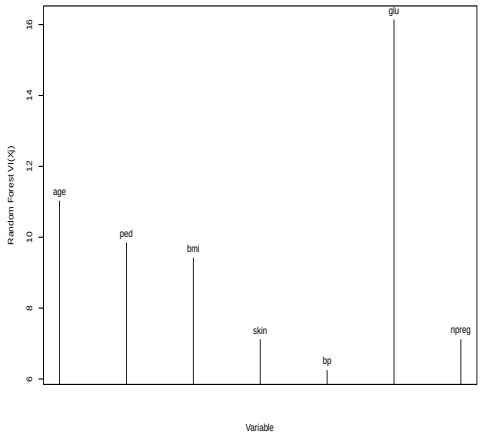



(h) Permutation-based Variable Importance.

**Fig. 4. Variable Importance Scores for the Attitude Data Set.** $n = 30$ **and** $p = 6$

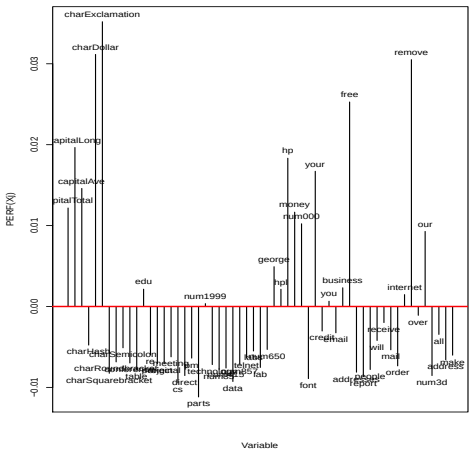



(i) Permutation-free Variable Importance.

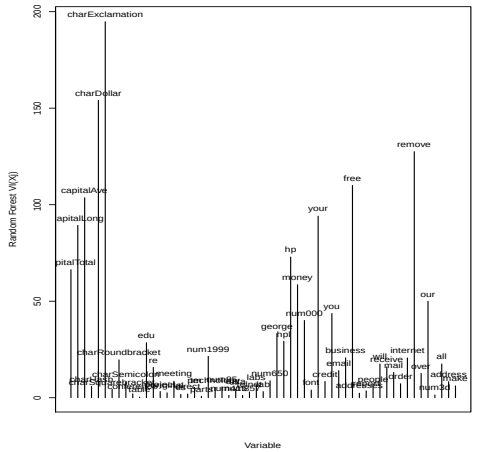



(j) Permutation-based Variable Importance.

**Fig. 5. Variable Importance Scores for the Spam Detection Dataset.** $n = 200$**,** $p = 7$**, and** $K = 2$ **classes**




(k) Permutation-free Variable Importance.




(l) Permutation-based Variable Importance.

**Fig. 6. Variable Importance Scores for the Spam Detection Dataset,** $n = 4601$**,** $p = 57$ **and** $K = 2$ **classes**

## Acknowledgements

Ernest Fokoué wishes to express his heartfelt gratitude and infinite thanks to Our Lady of Perpetual Help for Her ever-present support and guidance, especially for the uninterrupted flow of inspiration received through Her most powerful intercession.

## Competing Interests

Author has declared that no competing interests exist.

## References

[1] Breiman L. Random forests. Machine Learning. 2001;45:5-32.

[2] Breiman L. Statistical modeling: The two cultures. Statistical Science. 2001;16(3):199-215.

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

[4] Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics. 2008;9(307):1-12.

[5] Kuhn M. Building predictive models in r using the caret package. Journal of Statistical Software. 2008;28(5).

[6] Strobl C, Hothorn T, Zeileis A. Party on. R Journal. 2009;1(2):14-17.

[7] Shih YS. Regression trees with unbiased variable selection. Statistica Sinical. 2009;12:361-386.

[8] Sandri M, Zuccolotto P. A bias correction algorithm for the gini variable importance measure in classification trees. Journal of Computational and Graphical Statistics. 2008;17(3):611-628.



***Peer-review history:***
*The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)*
*http://sciencedomain.org/review-history/13781*