
DTU

“Deciphering Random Forest models through conditional variable importance”

Enbis presentation 2021

Marta Rotari[1], Murat Kulahci[1,2]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark
[2]Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden

Agenda

- Introduction
- Variable selection with Boruta Algorithm
- Variable importance and Ranking
- Extensions
  - Extension 1: Conditional importance
  - Extension 2: Backward Strategy
- Simulations and Results
- Conclusions

Introduction

- In enviroments such as manufacturing, the focus is on understanding the data-driven model rather than predictive modeling

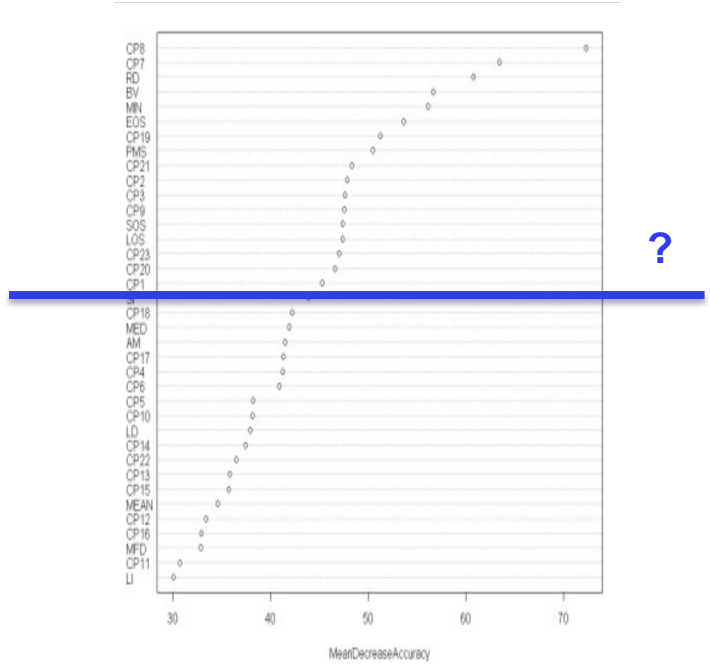
- Describe the contribution of the input variables to the model in the form of “variable importance”.
  Readily available in certain machine learning methods such as random forest (RF).

- Manufacturing data or more in general real world data are highly correlated.

- Research questions:

  - First problem: **cut-off point** or **threshold**. Which variables explain the output and have a greater impact on the response.

  - Second problem: **Correlated variables**. Does the correlation between variables have an effect on the ranking? If yes, how?



Variable selection

A Statistical Method for Determining Importance of Variables: **Boruta algorithm**.

The Boruta algorithm is a wrapper built around the random forest classification and regression algorithm. It is based on multiple application of RF and utilization of the estimate importance (type 1 = mean decrease in accuracy) generated by each RF.

It is a tool for estimating the importance of variables and its purpose is to capture all the important features with respect to an outcome variable.

→ **Outcomes**:

- Classification of all variables in two groups: Important and unimportant.
- Mean of the importance of all the variables among all the iterations.

Rudnicki WR, Kierczak M, Koronacki J, Komorowski "A Statistical Method for Determining Importance of Variables in an Information System."

Boruta algorithm

- Find the max score among shadow attributes (MZSA). Find every attribute that scored better than MZSA.
- Deem the attributes which have importance significantly higher than MZSA as ‘important’ and the attributes which have importance significantly lower than MZSA as ‘unimportant’
- Remove all shadow attributes.
- Repeat until the importance is assigned for all the attributes or the algorithm reach the limit of iterations.
- At the end, a t-test is conducted to test the mean equality of variables classified as important and MZSA mean importance. The null hypothesis is rejected at 0.001 significance level.

Variable importance and ranking in random forest

- The classical variable importance (type 1):

For any j-th attribute, in every tree in the forest count the number of correct oob-predictions, $i \in B^{(t)}$.

Randomly permute the values of j-th attribute in the oob-objects. Count the number of correct predictions.

Subtract the latter number of votes from the previous one.

$$I(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left(\sum_{i \in \mathbb{B}^{(t)}} I(y_i = \hat{y}_i^{(t)}) - \sum_{i \in \mathbb{B}^{(t)}} I(y_i = \hat{y}_{i,\pi_j}^{(t)}) \right)$$

x_1	x_2		x_j			x_n
			$x_{\pi(1)\mathrm{j}}$			
			$x_{\pi(2)\mathrm{j}}$			
			$x_{\pi(n)\mathrm{j}}$			

Correlated variables

Let us suppose that in our data set we have correlated variables:

What is the effect of correlation on the variable ranking of Random Forest?

The permutation importance is less able to detect the most relevant variables when the correlation increases.

Two key effects of the correlation on the permutation importance measure:

- the importance values of the most discriminant correlated variables are not necessarily higher than a less discriminant one
- the permutation importance measure depends on the size of the correlated groups.

Let us consider a variable of interest Y, a vector of random variables $X = (X_1, .., X_p)$ and the following regression setting:

$$Y = \sum_{j=1}^{p} f_j(X_j) + \epsilon$$

where ϵ is such that $E[\epsilon|X] = 0$, f_j are measurable functions so it can be decomposed as $f(x) = \sum_{j=1}^{p} f_j(x_j)$. For any $j \in \{1, .., p\}$ the permutation importance measure satisfies

$$I(X_j) = 2\mathbb{V}[f_j(X_j)].$$

In this framework, the permutation importance corresponds to the variance of $f_j(X_j)$, up to a factor 2.

Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre. "Correlation and variable importance in random forests." *Statistics and Computing* 27.3 (2017): 659-678.

Extension 1

We extend the Boruta algorithm for correlation data by using Conditional Importance.

Conditional importance:

In any tree in the forest count the number of correct oob-predictions.
Identify a subgroup Z of variable among all variables except X_j: $Z = X_1, X_2, \ldots, X_p$.
Within this grid permute the values of X_j and compute the oob-prediction accuracy after permutation.

The difference between the prediction accuracy before and after the permutation gives the importance of X_j for one tree. The average over all trees gives the X_j importance.

To determine the variables Z to be conditioned on, the most intuitive choice is to include all variables whose empirical correlation with X_j exceeds a certain threshold.

Another option is to let the user select certain variables to condition on, i.e., if a hypothesis of interest includes certain independencies

Y	X_j	Z
y_1	$x_{\pi_{j\|Z=a}(1),j}$	$z_1 = a$
y_3	$x_{\pi_{j\|Z=a}(3),j}$	$z_3 = a$
y_{27}	$x_{\pi_{j\|Z=a}(27),j}$	$z_{27} = a$
y_6	$x_{\pi_{j\|Z=b}(6),j}$	$z_6 = b$
y_{14}	$x_{\pi_{j\|Z=b}(14),j}$	$z_{14} = b$
y_{21}	$x_{\pi_{j\|Z=b}(21),j}$	$z_{21} = b$
$\vdots$	$\vdots$	$\vdots$

Extension 2

- We exploit the idea of the Boruta algorithm by Backward strategy.
- Using Recursive Feature Elimination
- Algorithm:
  - Extend the information system by adding some Shadow attributes.
  - Run Random Forest and use Conditional importance to rank the variables.
  - Compute the max score of the shadows variables.
  - Variables which systematically show importance score less than the noise variables are deemed as "Rejected" and removed from the dataset

1. First problem: cut-off point or threshold. Which variables keep and which ones to eliminate.

2. Second problem: Correlated variables. Does the correlation between variables have an effect on the ranking? If yes, how?

- Statistical Method for Determining Importance of Variables:

  Boruta alghoritm

- Conditional Variable Importance for Random Forest

Ex 1: Extended Boruta extended with the conditional importance for the correlation data sets

Ex 2: Backward elimination with conditional importance

Simulations and Results

- $\boldsymbol{X}$ is an artificial dataset composed of 20 variables drawn from joint standard normal distribution.

Two groups of correlation was introduced: x_1, x_2, x_3, x_4, x_5 and $x_{10}, x_{11}, x_{12}, x_{13}, x_{14}$

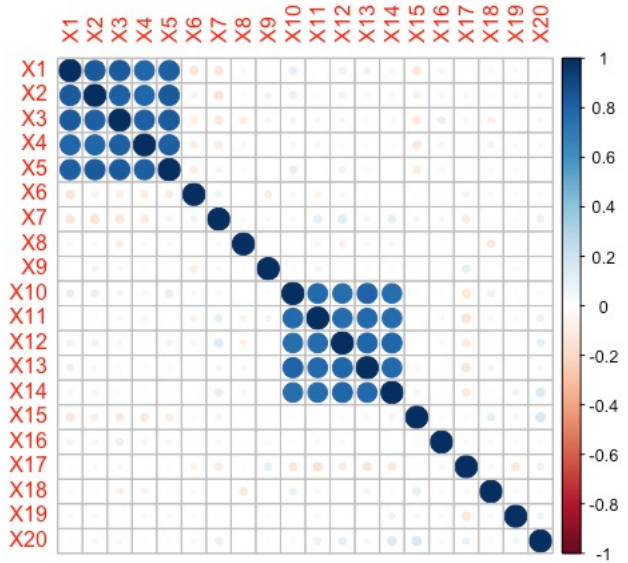
The Y is generated as a linear combination of X' : $Y = X'\beta + \epsilon$

where $X' = [X_2, X_{11}, X_{19}, X_{20}]$, β is such that $\frac{\beta}{sd(\beta)} > 2$ and $\epsilon \sim N(0,0.1)$.

Simulations:

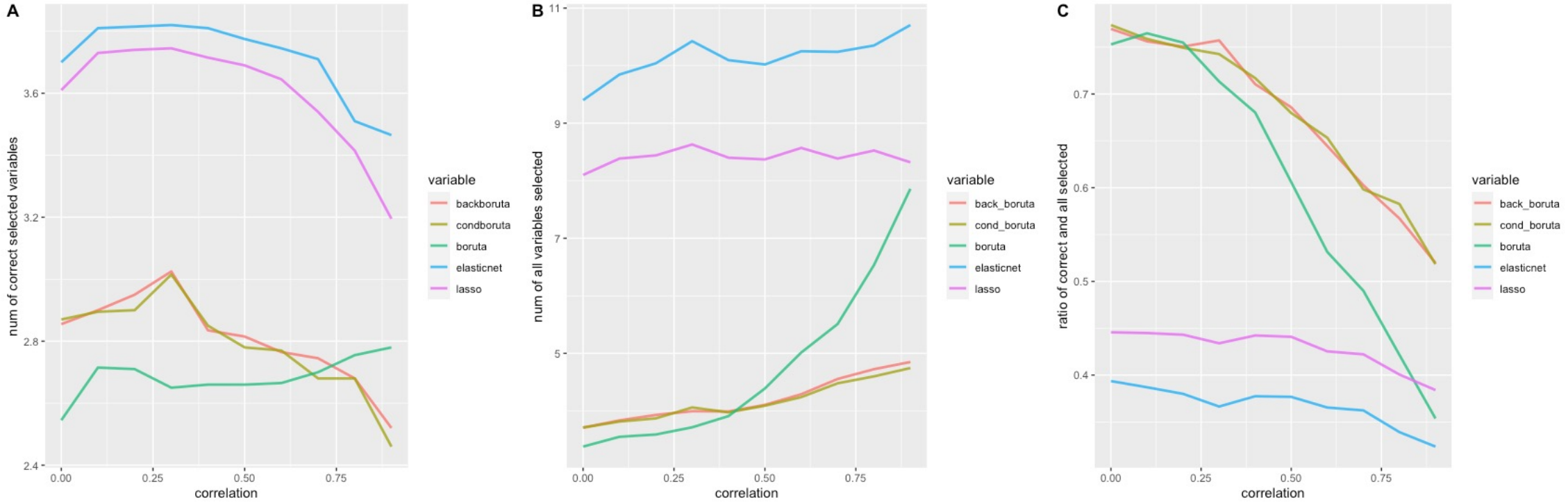
- Correlation [0,0.9] and we run 200 iterations

- Compared:
  - Boruta
  - Conditional boruta (extension 1)
  - Backward Boruta (extension 2)
  - Lasso
  - Elastic net



Results 1: Selected Variables

In our simulated data sets, the important variables are $[x_2, x_{11}, x_{19}, x_{20}]$.



A: correlation vs the number of correct selected variables.

B: correlation vs the number of all selected variables.

C: correlation vs the ratio between the correct selected and all selected variables.

Conclusions

- The correlation between data is crucial when doing feature selection.
- From the simulation results we observed that in the case of highly correlated data the two extensions perform better.
- Especially in the presence of high correlation Conditional Boruta and Backward Boruta give us a much more precise and accurate view of the important variables.
- This approach can be used in many industrial applications by providing more transparency and understanding of the process.

Thank your for your attention!

Marta Rotari
mrot@dtu.dk