



Random Forest: A Cautionary Note on Visualisations

Alan Inglis, Andrew Parnell, Catherine Hurley

Contents

- Introduction
- Random Forests
- Variable importance
 - Methods for calculating variable importance
 - Percentage increase in MSE
 - Increase in node purity
- Minimal Depth of a tree
 - Maximal subtrees
- Conditional Depth
- Misleading depth plots
- References

Introduction

- Random forests are machine learning models known for achieving strong predictive performance across a wide variety of domains.
- However, this strong performance comes at the cost of interpretability.
- A feature of RF is its ability to evaluate the relative importance of the predictor/explanatory variables.
- The aim of this presentation is to show different methods of displaying RF variable importance and examine some misleading properties of some of these visualisations.

What are Random Forests

- A decision tree is a supervised learning method used to create a predictive model for classification and regression.
- A Random Forest is an ensemble of decision trees.
- The principle of using an ensemble of decision trees is to achieve a more accurate prediction



Variable Importance

Variable importance is calculated in two ways in randomForest package.[1]

The importance measures show how much MSE/Impurity increase there is when that variable is randomly permuted.

Important variables will change the predictions by quite a bit if randomly permuted.

How is *Variable Importance* Calculated in randomForest package?

In the randomForest package, there are two measures of importance.

For Regression:

Percentage increase in the mean squared error (%IncMSE)

Increase in node purity, (IncNodePurity)

For Classification:

Mean decrease in accuracy

Mean decrease in the Gini index

%IncMSE

$$OOBMSE_t = \frac{1}{n_{OOB_t}} \sum_{i \in OOB_t}^n (y_i - \tilde{y}_{i,t})^2$$

$$OOBMSE_t(X_j) = \frac{1}{n_{OOB_t}} \sum_{i \in OOB_t}^n (y_i - \tilde{y}_{i,t}(X_j))^2$$

$$OOBMSE_t(X_j) - OOBMSE_t$$

- The $OOBMSE_t$ is calculated for a tree t .
- The data is permuted.
- The MSE is calculated again for $OOBMSE_t(X_j)$ (where X_j is the permuted value of variable X_j)
- Difference between the two are then averaged over all trees.

%IncMSE

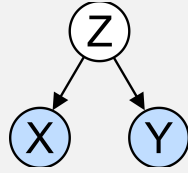
The idea behind this variable importance measure is that if the variable X_j is an important predictor of the response, then permuting the values of that variable should lead to a sharp increase in residual sum of squares.[3]

But if the variable is unimportant to the prediction of the response variable, then permuting its values should have little effect on the predictions.

IncNodePurity

- This measure is the total decrease in node impurities from splitting on the variable.
- It is averaged over all nodes in which the variable is split and over all trees.
- It is analogous to Gini-based importance & is calculated based on the reduction in the sum of squared errors whenever a variable is chosen to split.

Minimal Depth



Variables used for splitting close to the root tend to be important [2][3]



Maximal subtrees can be a practical tool in exploring random forests



Maximal subtrees look at how close they are to the root of a tree and help assess the importance of a given variable



This approach aims at exploring the structure of the forest instead of treating it like a black box

Maximal Subtree Example

- For each predictor X_j , we call Tx_j an X_j - *subtree* of our tree T , if the root of Tx_j is split using X_j .
- Tx_j is a *maximal* X_j - *subtree* if it is not a subtree of a larger Tx_j - *subtree*.
- In Figure 1.0, the *maximal* X_1 - *subtrees* are coloured in red.

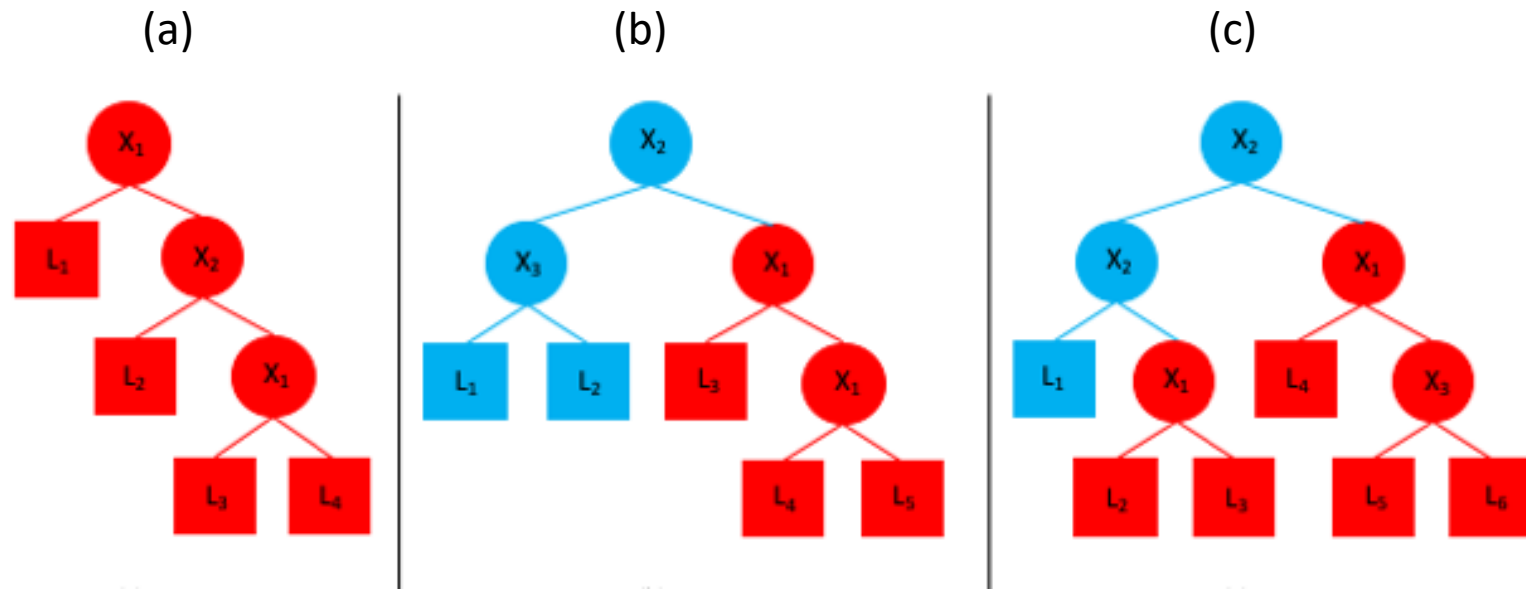


Figure 1.0: Illustration of maximal subtrees

- As X_1 splits at the root, the whole tree is a *maximal* X_1 - *subtree*.
- The *maximal* X_1 - *subtree* contains a X_1 - *subtree* that is not maximal.
- There are 2 *maximal* X_1 - *subtrees*, of which, one is closer to the root than the other.

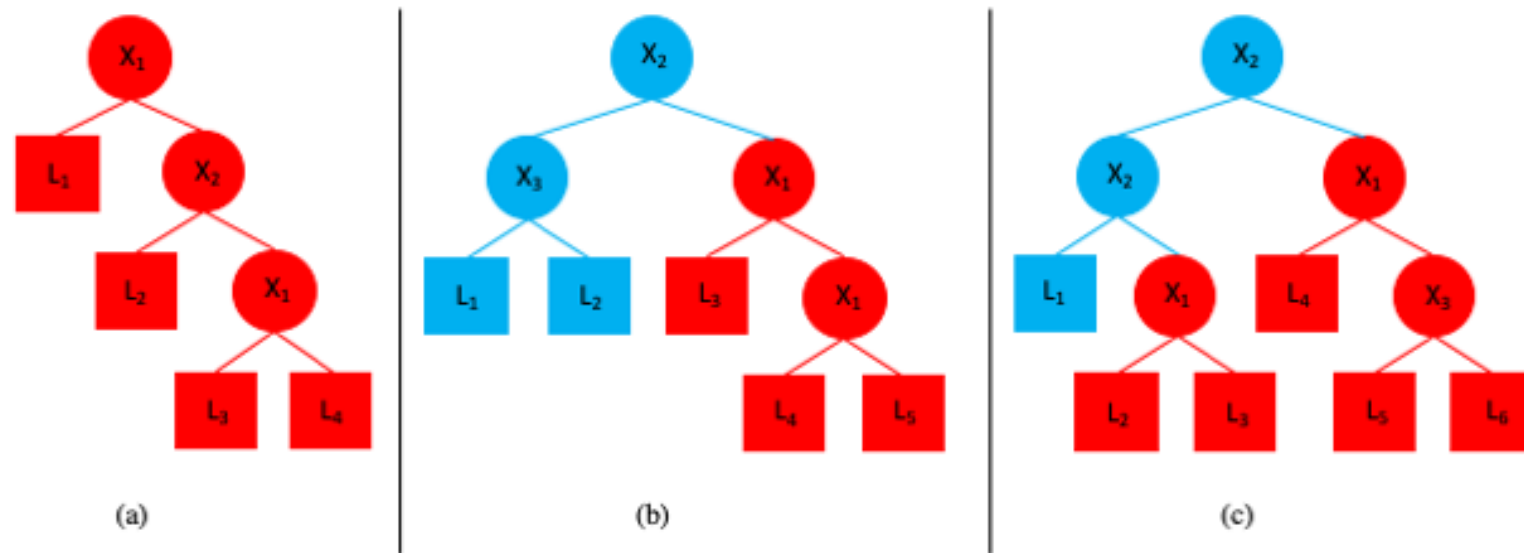
Conditional Minimal Depth

Conditional minimal depth is a modification of minimal depth.


Conditional minimal depth adjusts the concept of importance of single variables such that it would allow for interactions.

For a conditioning variable X_j and a second variable X_l , we define the conditional minimal depth of X_l with respect to X_j as the minimal depth of X_l in the maximal X_j -subtree closest to the root of the whole tree.[4]

Conditional Minimal Depth



- For example, in trees (b) & (c).
- X_3 is always below X_2 .
- So, conditional minimal depth of X_3 with respect to X_2 is equal to 1 and 2, respectively.
- On the other hand, the conditional minimal depth of X_2 with respect to X_3 is equal to the mean depth of maximal X_3 -subtrees in the forest as there are no splits on X_2 in X_3 -subtrees. [4]



Simulation Study

Several different models were simulated

Ranging from basic linear regression to more complex models that contain variable interactions

A random forest was fitted for each model

The variable importance and the minimal depth/minimal depth interactions were examined

Simulated Data

Standard Linear regression:

$$y = 10 + 10x_1 - 10x_2 + 10x_3 + \varepsilon$$

$$x_i \sim N(0,1), \varepsilon \sim N(0,1)$$

Interaction model 1:

$$y = 10 + 10x_1 + 10x_2 + 5x_1x_2 + \varepsilon$$

$$x_2 \sim N(0,1), x_1 \sim \text{bernoulli}(0.5) \\ \varepsilon \sim N(0,1)$$

Interaction model 2:

$$y = 10 + 10x_1 - 10x_1z_1 + \varepsilon$$

$$x \sim N(0,1), z \sim \text{bernoulli}(0.5), \\ \varepsilon \sim N(0,1)$$

Standard Linear Regression

$$y = 10 + 10x_1 - 10x_2 + 10x_3 + \varepsilon$$

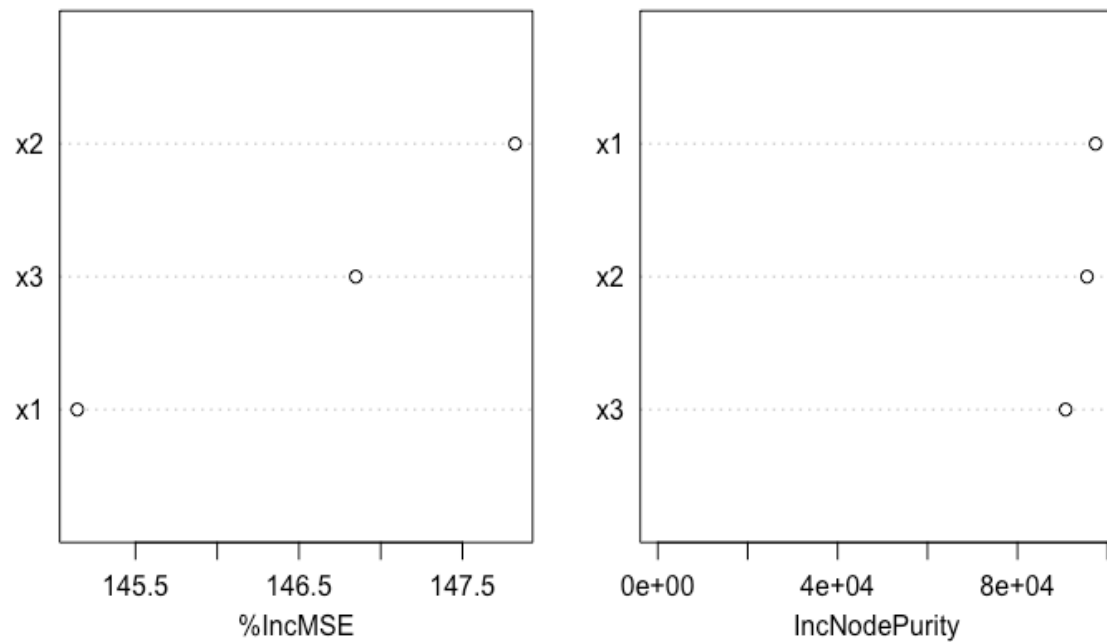


Figure 3.0: Variable importance plot

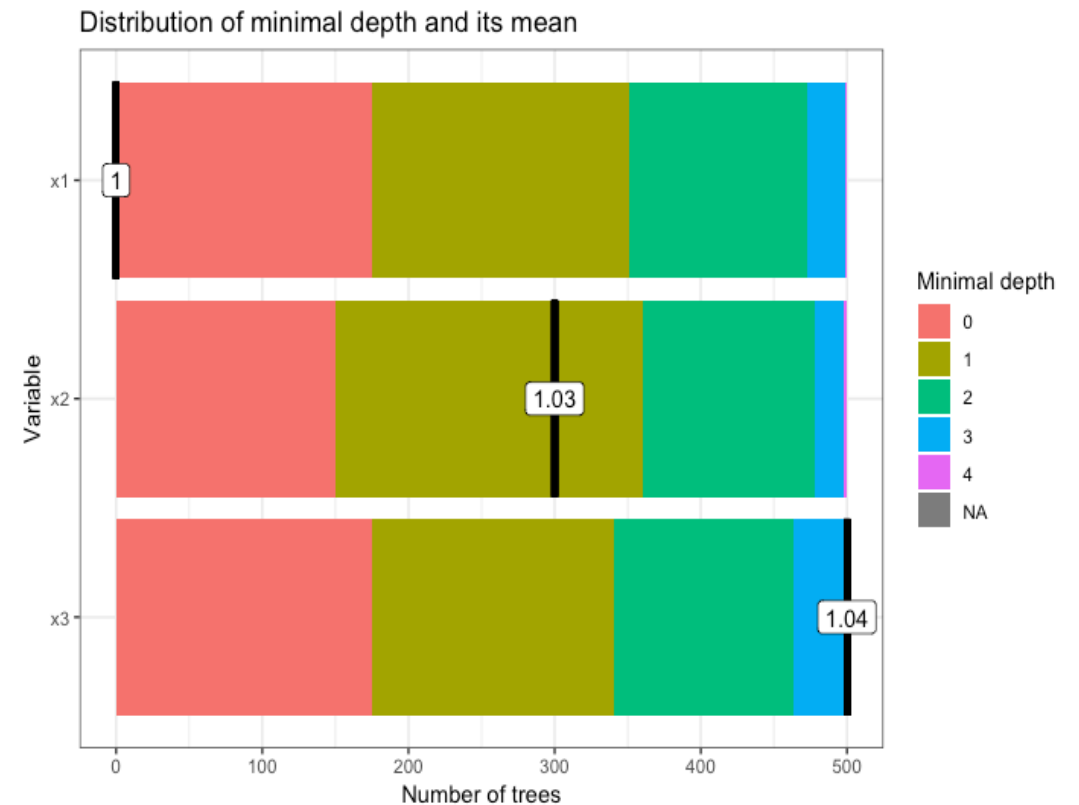


Figure 3.1: Minimal depth plot

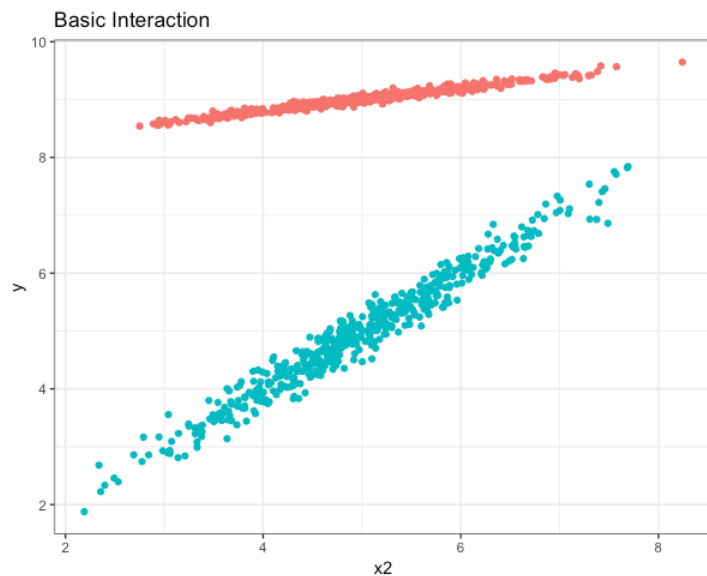


Figure 4.0: Interaction_1 plot

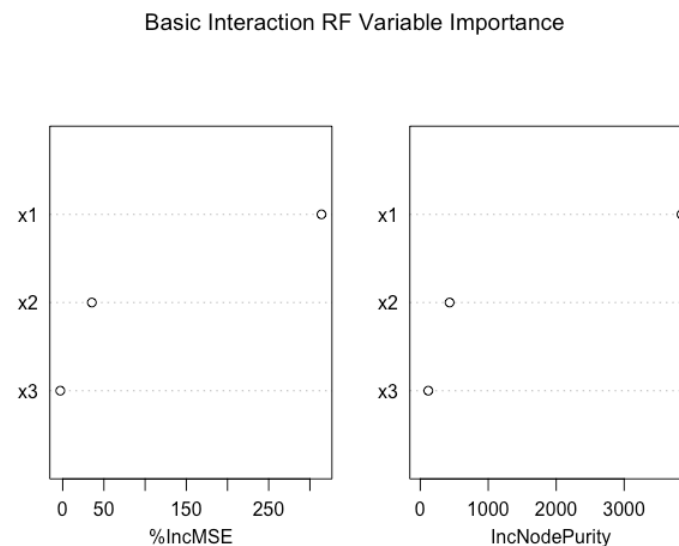


Figure 4.1: Variable importance plot

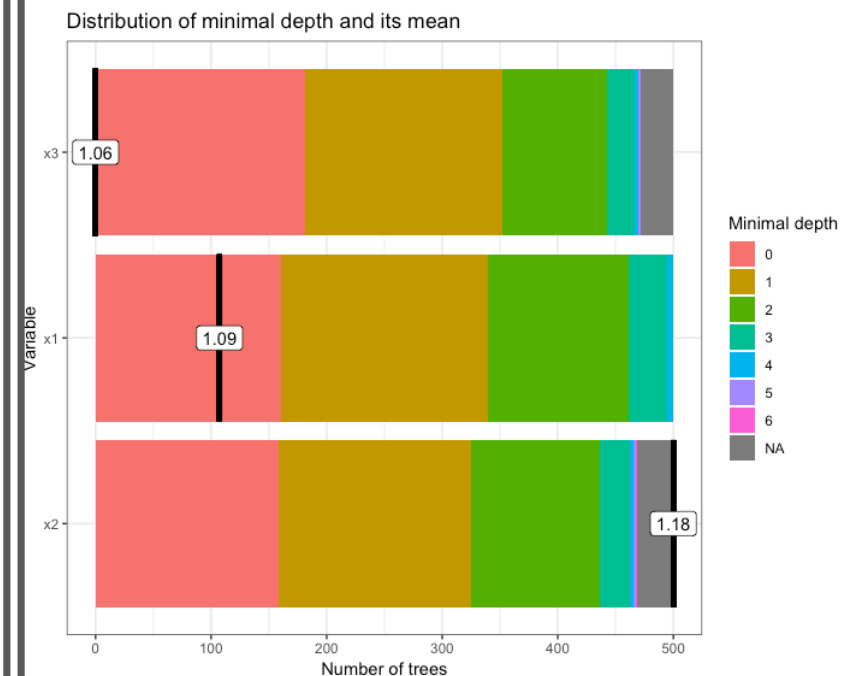


Figure 4.2: Minimal depth plot

Interaction model 1

$$y = 10 + 10x_1 + 10x_2 + 5x_1x_2 + \varepsilon$$

Interaction model 1

$$y = 10 + 10x_1 + 10x_2 + 5x_1x_2 + \varepsilon$$

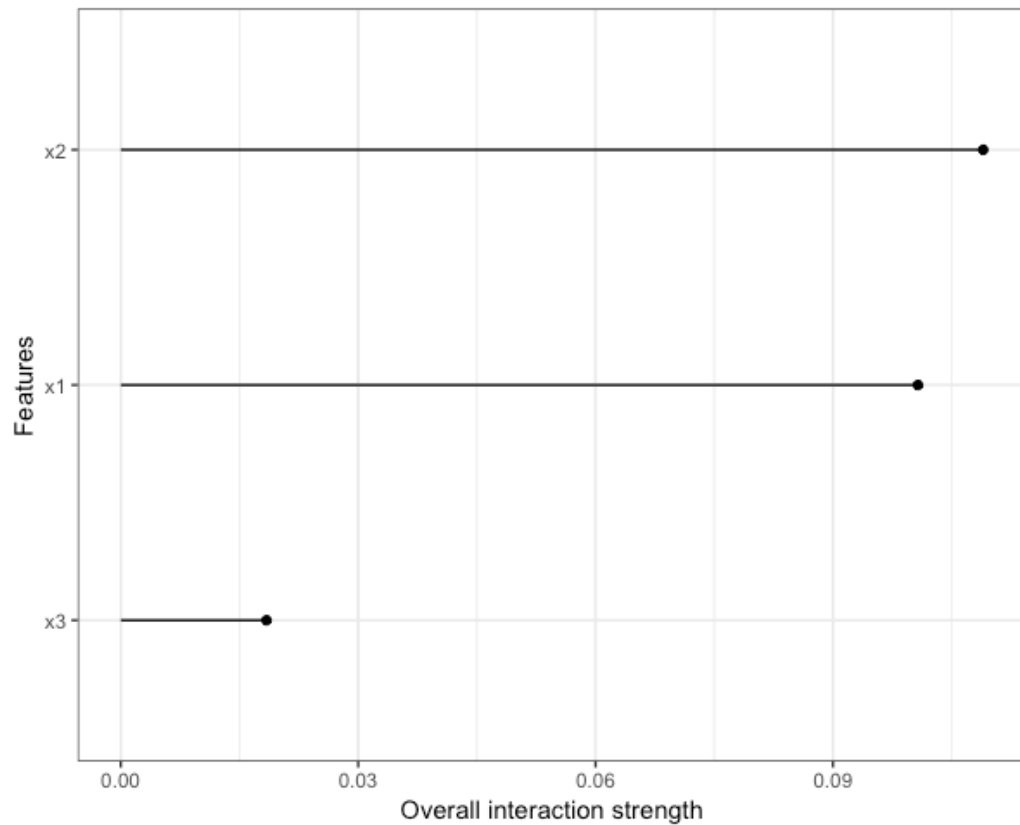


Figure 5.0: Overall interaction strength

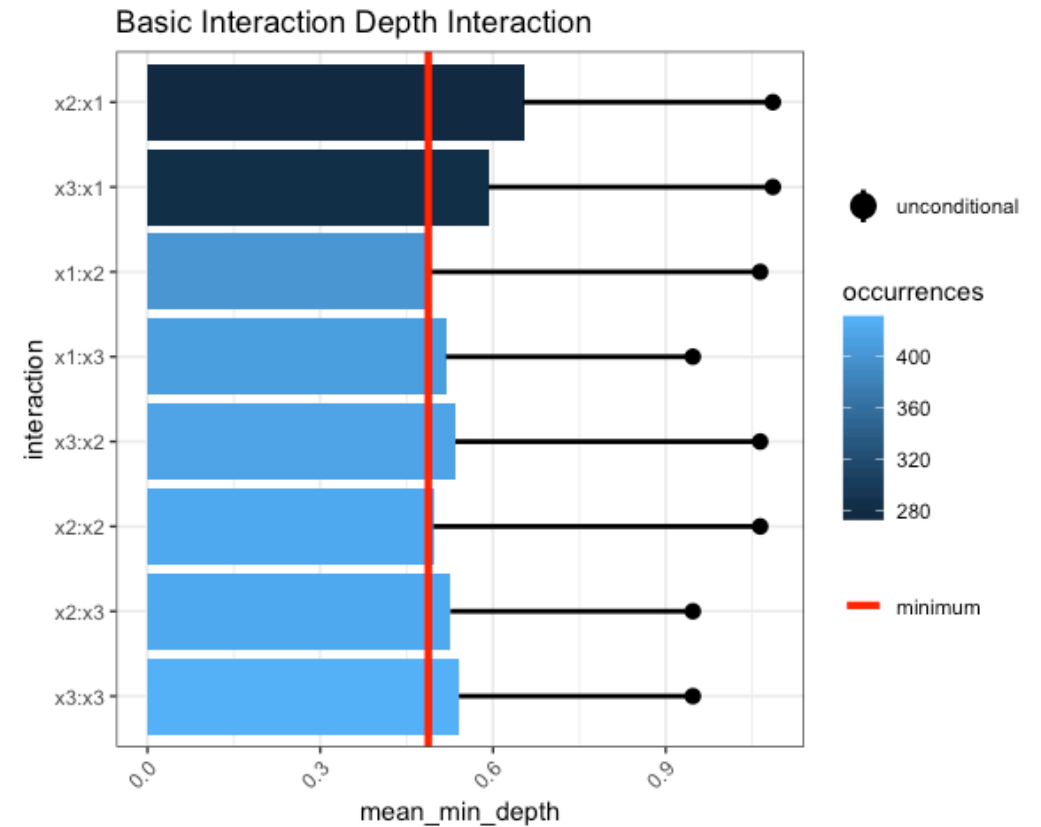


Figure 5.1: Minimal depth interaction

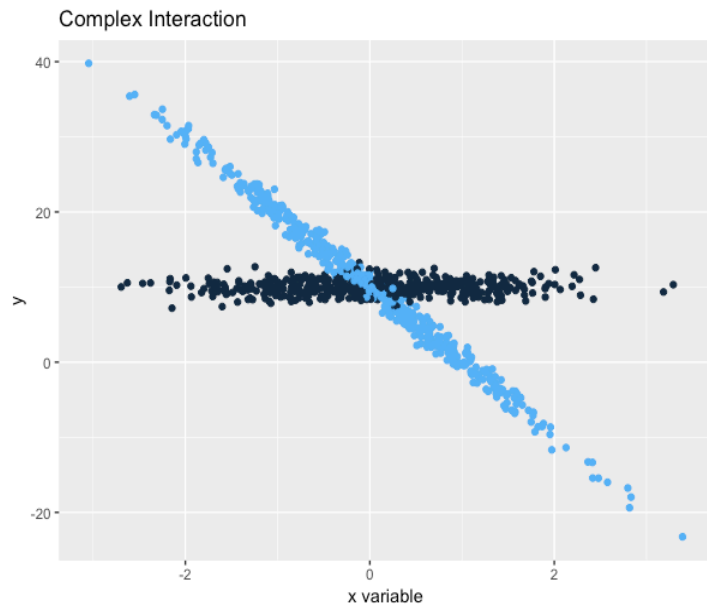


Figure 6.0: Interaction_2 plot

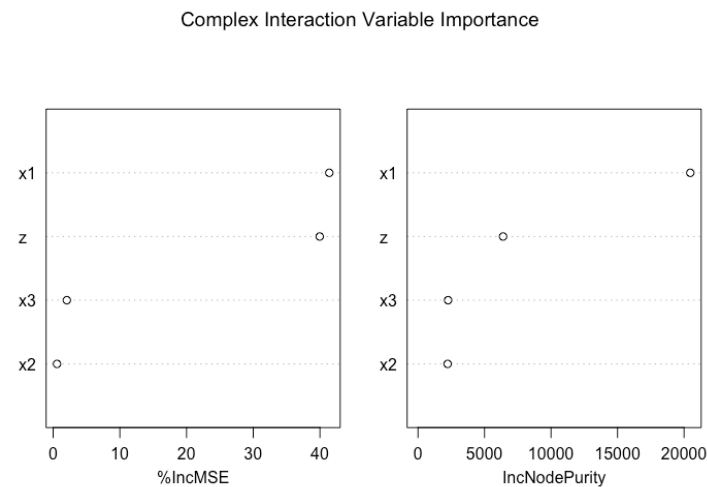


Figure 6.1: Variable importance plot

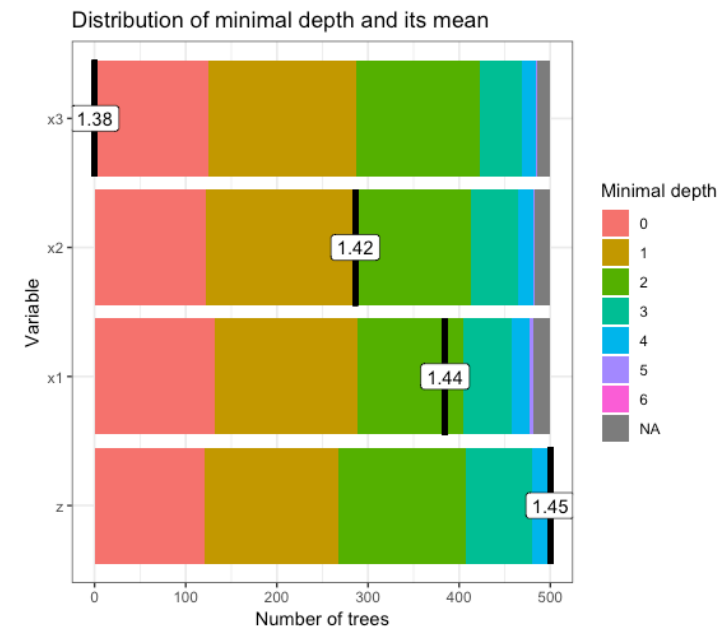


Figure 6.2: Minimal depth plot

Interaction model 2

$$y = 10 + 10x_1 - 10x_1z_1 + \varepsilon$$

Interaction model 2

$$y = 10 + 10x_1 - 10x_1z_1 + \varepsilon$$

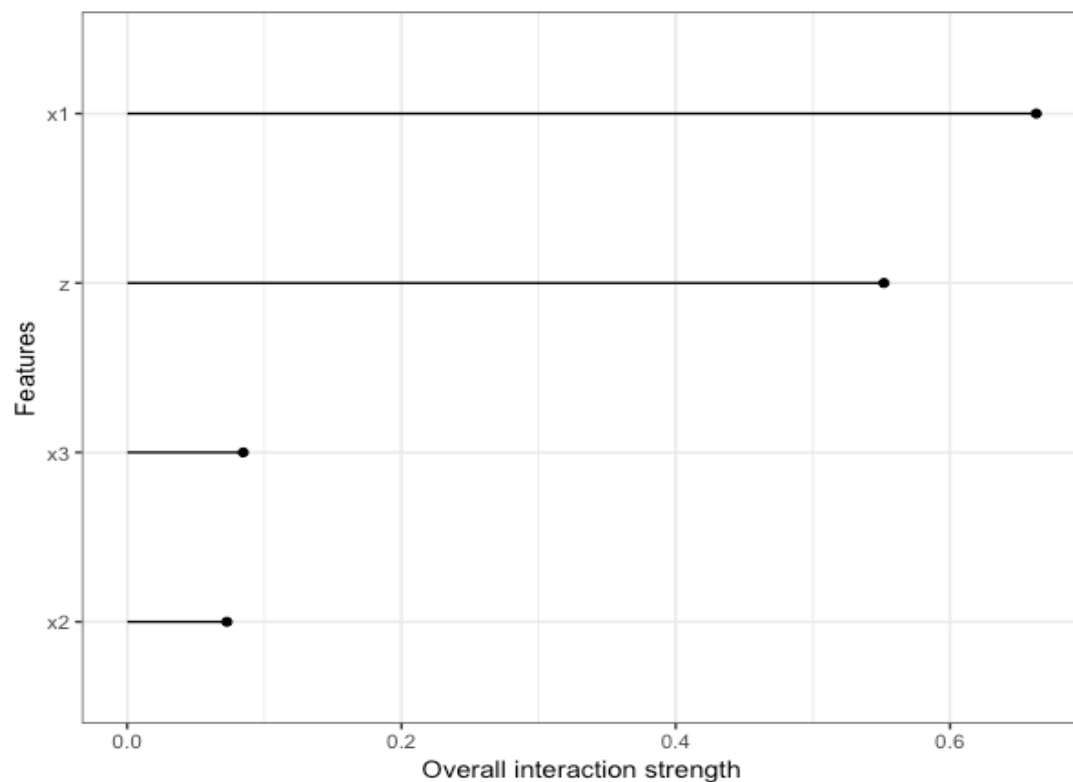


Figure 7.0: Overall interaction strength

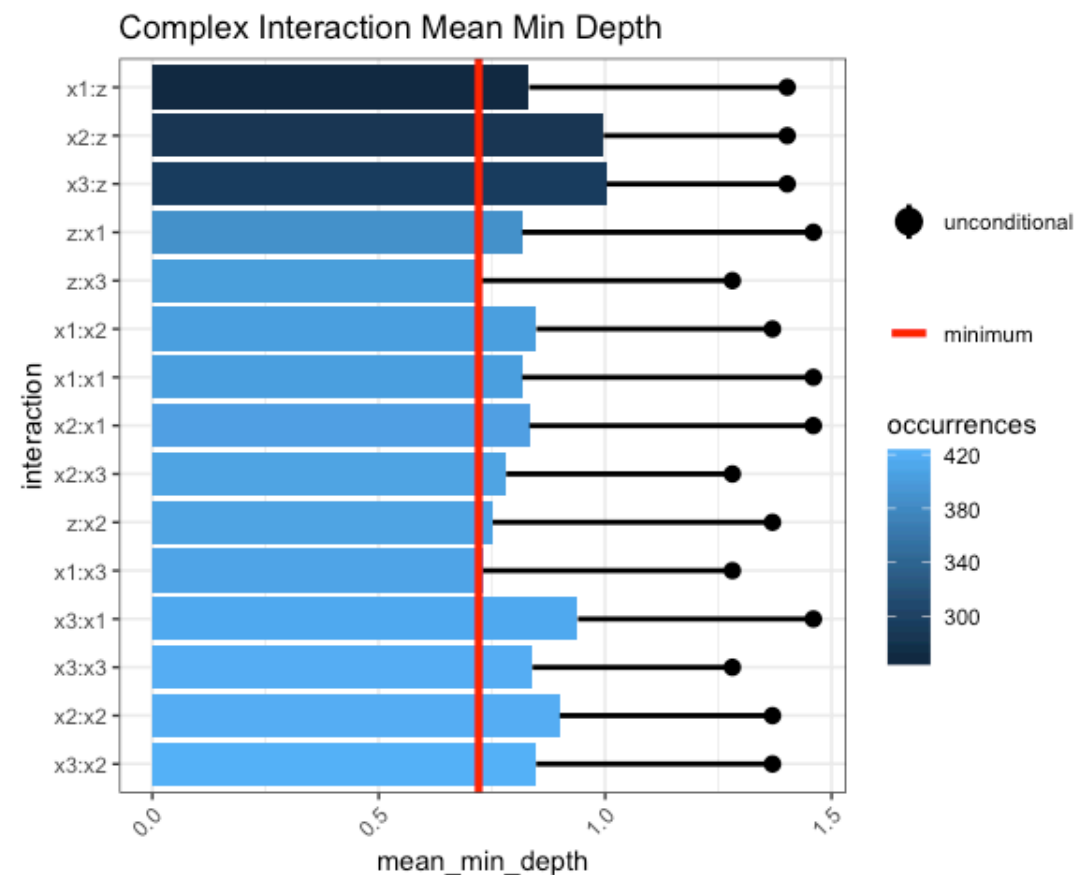


Figure 7.1: Minimal depth interaction

Mean Minimal Depth

- The mean minimal depth for a strong variable is smaller than a weak variable as long as p is not too large. i.e., $p < 1000$.
- In this range, minimal depth thresholding is highly effective.
- However, as p increases, the tree becomes overwhelmed with variables, and eventually the distribution degenerates to $D(T)$, and variable selection is no longer effective.

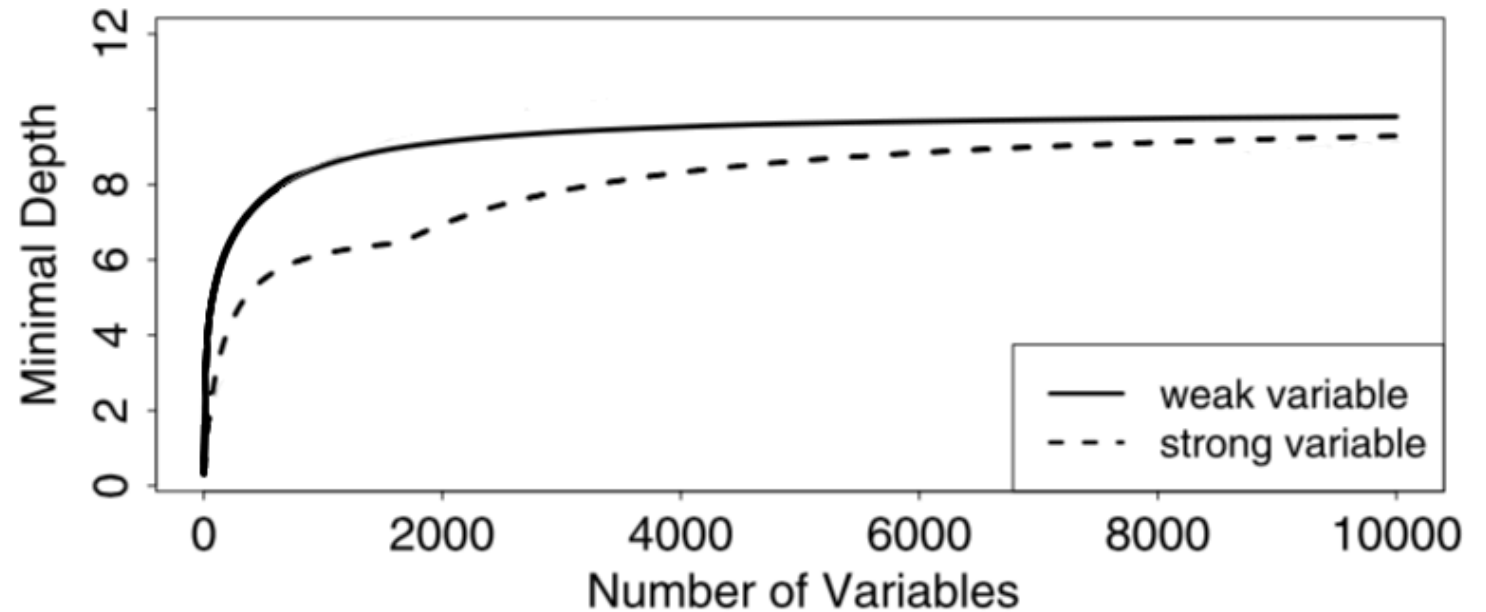


Fig 8: The black lines show mean of the minimal depth for weak and strong variables assuming a tree with depth $D(T) = 10$. For small p , strong variables have smaller minimal depths, but as p increases, the minimal depth converges to $D(T)$ for both types of variables.

Summary

Model	Variable Importance (rF)	Minimal Depth (rFE)	Interactions (rFE)
Interaction model 1	✓	✗	✓
Interaction model 2	✓	✗	✗

Conclusion/Future Work

- randomForest package variable importance measures seemed to do a good job.
- Maximal subtrees are dimensionless and are not constrained by any specific measure of prediction error.
- Also, maximal subtrees, unlike other VIMP methods, can be studied in detail.
- Both the minimal depth/interactions struggled in finding the important variables.
- randomForestExplainer package, somewhat, fails at its intended purpose (i.e., the plots are hard to interpret).
- Compare with some other packages that look at variable importance and measure depth (e.g., VSURF, rangeR).

References:

- [1] A. Liaw and M. Wiener. Classification and regression by randomforest. R News, 2(3):18–22, 2002 .
- [2] Hemant Ishwaran, Udaya B. Kogalur, Eiran Z. Gorodeski, Andy J. Minn, and Michael S. Lauer. High-dimensional variable selection for survival data. Journal of the American Statistical Association, 105(489):205-217, 2010.
- [3] H. Ishwaran. Variable importance in binary regression trees and forests. Electronic Journal of Statistics,, 1:519–37, 2007.
- [4] A. Paluszynska and P. Biecek. randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance, 2017. R package version 0.9

