

# randomForestExplainer

## What's in the forest?

Aleksandra Paluszynska [aut, cre]  
Przemysław Biecek [aut, ths]  
University of Warsaw



### Basics

The aim of the **randomForestExplainer** package is to support structure exploration and visualisation for a random forest model.

Once you have a model created with the **randomForest** package, use following functions to examine its structure.

```
library(randomForest)
library(randomForestExplainer)

forest <- randomForest(PV1MATH~.,
data = pisa2015, localImp = TRUE)
```

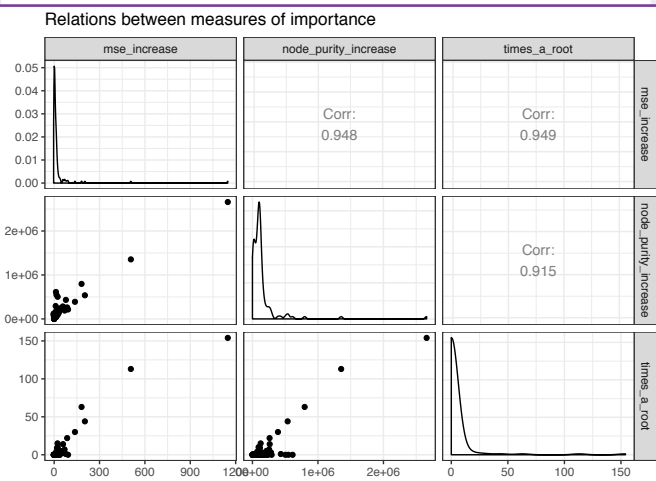
### Variable Importance

The **measure\_importance()** function calculates different measures of importance for variables presented in the forest. Note that different variables are available for classification forests and regression forests.

Use the **plot\_importance\_ggpairs()** function to plot examine relations between selected measures.

```
forest_stats <-
measure_importance(forest, measures =
c("mse_increase",
"node_purity_increase",
"times_a_root"))

plot_importance_ggpairs(forest_stats)
```



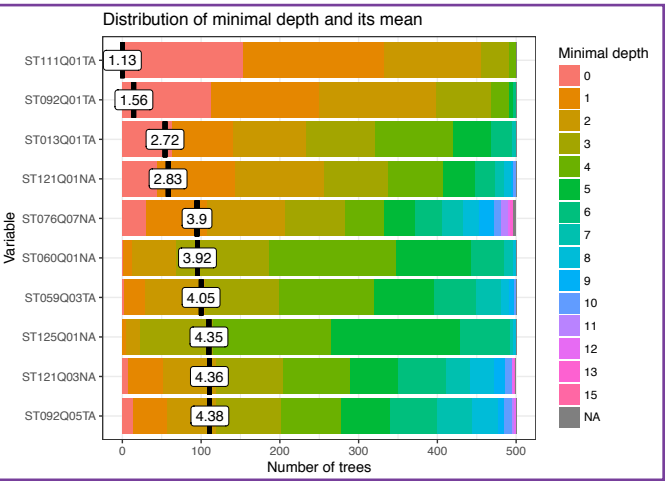
## randomForestExplainer - Structure mining and visualisation for Random Forests

### Variable Depth

The **min\_depth\_distribution()** function calculates distribution of minimal depth of given variable in all trees. Use the **plot\_min\_depth\_distribution()** function to plot this distribution along with mean depths for variables. In general, the higher are variables the more influential they are.

```
forest_frame <-
min_depth_distribution(forest)

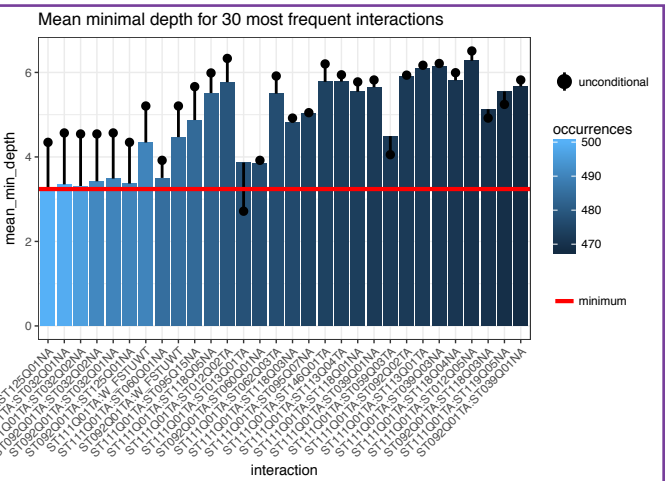
plot_min_depth_distribution(forest_frame)
```



The **min\_depth\_interactions()** function calculates conditional depth of variables in subtrees rooted in the selected variable. Such statistic is useful to identify interactions of two variables.

Use the **plot\_min\_depth\_interactions()** function to plot such statistics.

```
forest_interactions <-
min_depth_interactions(forest)
plot_min_depth_interactions(forest_interactions)
```

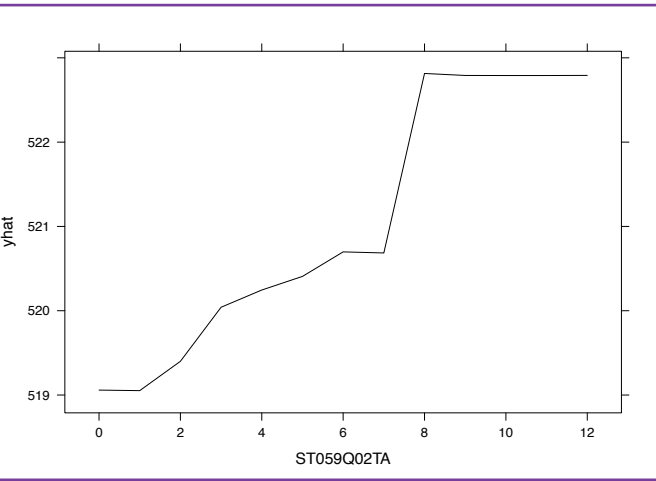


### Partial Dependence Plots

The **partial()** function from the **pdp** package calculates marginal relation between target variable and selected one or two independent variables. The relation can be noted with **lattice** graphical system with the **plotParial()** function or with **ggplot2** system with the **autoplot()** function.

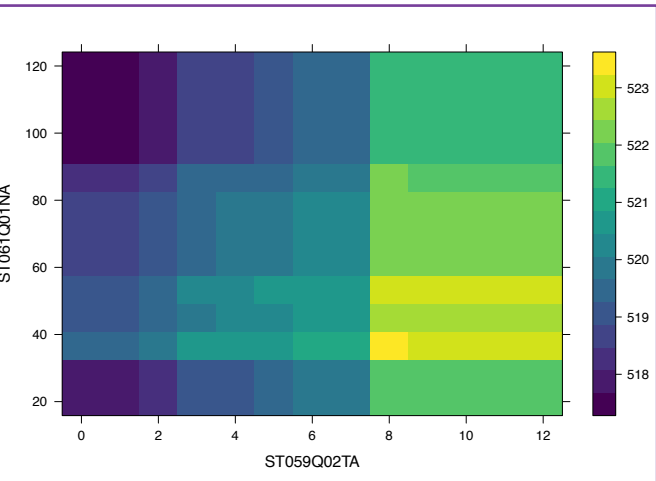
```
library(pdp)

pdp1 <- partial(forest, "ST059Q02TA")
plotPartial(pdp1)
```



Variable depth and variable importance functions are useful in identification which variable/variables are worth watching, while the **pdp** plots are useful to understand the nature of the relation between target variable and variable/s of interest.

```
pdp2 <- partial(forest, c("ST059Q02TA",
"ST061Q01NA"))
plotPartial(pdp2)
```



### Local Approximations

Based on LIME

### Literature

**ggRandomForests**: Random Forests for Regression. John Ehrlinger (2016)

**pdp**: An R Package for Constructing Partial Dependence Plots. Brandon M. Greenwell (2017)

**forestFloor**: Forest Floor Visualizations of Random Forests. Soeren Welling, Hanne Refsgaard, Per Brockhoff, Line Clemmensen (2016)