Package ‘randomForestVIP’

July 19, 2023

Type Package
Title Tune Random Forests Based on Variable Importance & Plot Results
Version 0.1.3
Description Functions for assessing variable relations and associations prior to modeling with a Random Forest algorithm (although these are relevant for any predictive model). Metrics such as partial correlations and variance inflation factors are tabulated as well as plotted for the user. A function is available for tuning the main Random Forest hyper-parameter based on model performance and variable importance metrics. This grid-search technique provides tables and plots showing the effect of the main hyper-parameter on each of the assessment metrics. It also returns each of the evaluated models to the user. The package also provides superior variable importance plots for individual models. All of the plots are developed so that the user has the ability to edit and improve further upon the plots. Derivations and methodology are described in Bladen (2022) <https://digitalcommons.usu.edu/etd/8587/>.
License GPL-3
URL https://github.com/KelvynBladen/randomForestVIP
Depends R (>= 4.0.0)
Imports car, dplyr, ggplot2, gridExtra, minerva, randomForest, stats, tidyr
Suggests EZtune, e1071, knitr, MASS, rmarkdown, rpart, testthat (>= 3.0.0)
VignetteBuilder knitr
Config/testthat/edition 3
Encoding UTF-8
LazyData true
RoxygenNote 7.2.3
NeedsCompilation no
Author Kelvyn Bladen [aut, cre], D. Richard Cutler [aut]
**boston**

**Maintainer** Kelvyn Bladen <kelvyn.bladen@usu.edu>

**Repository** CRAN

**Date/Publication** 2023-07-19 11:20:02 UTC

## R topics documented:

- **boston** (2)
- **ggvip** (3)
- **lichen** (4)
- **mtry_compare** (6)
- **partial_cor** (7)
- **robust_vifs** (7)

## Index

- **boston**

---

**Description**

The Boston data frame has 506 rows and 14 columns.

**Usage**

boston

**Format**

This data frame contains the following columns:

- **crim** per capita crime rate by town.
- **zn** proportion of residential land zoned for lots over 25,000 sq.ft.
- **indus** proportion of non-retail business acres per town.
- **chas** Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- **nox** nitrogen oxides concentration (parts per 10 million).
- **rm** average number of rooms per dwelling.
- **age** proportion of owner-occupied units built prior to 1940.
- **dis** weighted mean of distances to five Boston employment centres.
- **rad** index of accessibility to radial highways.
- **tax** full-value property-tax rate per $10,000.
- **ptratio** pupil-teacher ratio by town.
- **black** $1000(Bk - 0.63)^2$ where $Bk$ is the proportion of blacks by town.
- **lstat** lower status of the population (percent).
- **medv** median value of owner-occupied homes in $1000$s.
ggvip

Source
https://www.stats.ox.ac.uk/pub/MASS4/

---

ggvip Variable Importance GGPlot

Description
A ggplot of variable importance as measured by a Random Forest.

Usage
ggvip(x, scale = FALSE, sqrt = TRUE, type = "both", num_var)

Arguments
x
  An object of class randomForest.
scale
  For permutation based measures such as MSE or Accuracy, should the measures be divided by their "standard errors"? Default is False.
sqrt
  Boolean value indicating whether importance metrics should be adjusted via a square root transformation. Default is True.
type
  either 1 or 2, specifying the type of importance measure (1=mean decrease in accuracy or node impurity or mean decrease in gini). Default is "both".
um_var
  Optional argument for reducing the number of variables to the top 'num_var'. Must be an integer between 1 and the total number of predictor variables in the model.

Value
A ggplot dotchart showing the importance of the variables that were plotted.

Examples
rf <- randomForest::randomForest(factor(Species) ~ ., importance = TRUE, data = iris)
ggvip(rf, scale = FALSE, sqrt = TRUE, type = "both")
Lichen data from the Current Vegetation Survey

Description

Data were collected between 1993 and 1999 as part of the Lichen Air Quality surveys on public lands in Oregon and southern Washington. Observations were obtained from 1-acre (0.4 ha) plots at Current Vegetation Survey (CVS) sites. Indicator variables denote the presences and absences of 7 lichen species. Data for each sampled plot include the topographic variables elevation, aspect, and slope; bioclimatic predictors including maximum, minimum, daily, and average temperatures, relative humidity precipitation, evapotranspiration, and vapor pressure; and vegetation variables including the average age of the dominant conifer and percent conifer cover. The data in lichenTest were collected from half-acre plots at CVS sites in the same geographical region and contains many of the same variables, including presences and absences for the 7 lichen species. As such, it is a good test dataset for predictive methods applied to the Lichen Air Quality data.

Usage

lichen

Format

A data frame with 840 observations and 40 variables. One variable is a location identifier, 7 (coded as 0 and 1) identify the presence or absence of a type of lichen species, and 32 are characteristics of the survey site where the data were collected.

There were 12 monthly values in the original data for each of the bioclimatic predictors. Principal components analyses suggested that for each of these predictors 2 principal components explained the vast majority (95.0%-99.5%) of the total variability. Based on these analyses, indices were created for each set of bioclimatic predictors. The variables with the suffix Ave in the variable name are the average of 12 monthly variables. The variables with the suffix Diff are contrasts between the sum of the April-September monthly values and the sum of the October-December and January-March monthly values, divided by 12. Roughly speaking, these are summer-to-winter contrasts.

The variables are summarized as follows:

- LobaOreg: Lobaria oregana (Absent = 0, Present = 1)
- EvapoTransAve: Average monthly potential evapotranspiration in mm
- EvapoTransDiff: Summer-to-winter difference in monthly potential evapotranspiration in mm
- MoistIndexAve: Average monthly moisture index in cm
- MoistIndexDiff: Summer-to-winter difference in monthly moisture index in cm
- PrecipAve: Average monthly precipitation in cm
- PrecipDiff: Summer-to-winter difference in monthly precipitation in cm
- RelHumidAve: Average monthly relative humidity in percent
- RelHumidDiff: Summer-to-winter difference in monthly relative humidity in percent
- PotGlobRadAve: Average monthly potential global radiation in kJ
PotGlobRadDiff  Summer-to-winter difference in monthly potential global radiation in kJ
AveTempAve   Average monthly average temperature in degrees Celsius
AveTempDiff  Summer-to-winter difference in monthly average temperature in degrees Celsius
MaxTempAve   Average monthly maximum temperature in degrees Celsius
MaxTempDiff  Summer-to-winter difference in monthly maximum temperature in degrees Celsius
MinTempAve   Average monthly minimum temperature in degrees Celsius
MinTempDiff  Summer-to-winter difference in monthly minimum temperature in degrees Celsius
DayTempAve   Mean average daytime temperature in degrees Celsius
DayTempDiff  Summer-to-winter difference in average daytime temperature in degrees Celsius
AmbVapPressAve Average monthly average ambient vapor pressure in Pa
AmbVapPressDiff Summer-to-winter difference in monthly average ambient vapor pressure in Pa
SatVapPressAve Average monthly average saturated vapor pressure in Pa
SatVapPressDiff Summer-to-winter difference in monthly average saturated vapor pressure in Pa
Aspect   Aspect in degrees
TransAspect Transformed Aspect: TransAspect=(1-cos(Aspect))/2
Elevation   Elevation in meters
Slope   Percent slope
ReserveStatus Reserve Status (Reserve, Matrix)
StandAgeClass Stand Age Class (< 80 years, 80+ years)
ACONIF   Average age of the dominant conifer in years
PctVegCov  Percent vegetation cover
PctConifCov Percent conifer cover
PctBroadLeafCov Percent broadleaf cover
TreeBiomass Live tree (> 1 inch DBH) biomass, above ground, dry weight

Source
https://CRAN.R-project.org/package=EZtune/
mtry_compare  

Mtry Tune via VIPs

Description
A list of data.frames and useful plots for user evaluations of the randomForest hyperparameter mtry.

Usage

mtry_compare(
  formula,
  data = NULL,
  scale = FALSE,
  sqrt = TRUE,
  num_var,
  mvec,
  ...
)

Arguments

formula  an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted.
data  an optional data frame containing the variables in the model. By default the variables are taken from the environment which randomForest is called from.
scale  For permutation based measures such as MSE or Accuracy, should the measures be divided by their "standard errors"? Default is False.
sqrt  Boolean value indicating whether importance metrics should be adjusted via a square root transformation. Default is True.
num_var  Optional integer argument for reducing the number of variables to the top 'num_var'. Should be an integer between 1 and the total number of predictor variables in the model or it should be a positive proportion of variables desired.
mvec  Optional vector argument for defining choices of mtry to have the function consider. Should be a vector of integers between 1 and the total number of predictor variables in the model. Or it can be a vector of proportions (strictly less than 1) of the number of predictor variables.
...
Other parameters to pass to the randomForest function.

Value
A list of data.frames, useful plots, and forest objects for user evaluations of the randomForest hyperparameter mtry.

Examples

m <- mtry_compare(factor(Species) ~ ., data = iris, sqrt = TRUE)
m
partial_cor

Partial Correlations

Description
A list of data.frames and useful plots for user evaluations of correlations and partial correlations of
predictors with a given response.

Usage
partial_cor(formula, data = NULL, model = lm, num_var, ...)

Arguments
- formula: an object of class "formula" (or one that can be coerced to that class): a symbolic
description of the model to be fitted.
- data: a data frame containing the variables in the model. By default the variables are
taken from the environment which the model is called from.
- model: Model to use for extraction partial correlations. Possible model choices are lm,
rpart, randomForest, and svm. Default is lm.
- num_var: Optional integer argument for reducing the number of variables to the top 'num_var'.
  Should be an integer between 1 and the total number of predictor variables in
  the model or it should be a positive proportion of variables desired.
- ...: Additional arguments to be passed to model as needed.

Value
A list of data.frames and useful plots for user evaluations of partial correlations.

Examples
pcs <- partial_cor(Petal.Length ~ ., data = iris[-5], model = lm)
pcs$plot_y_part_cors

robust_vifs

Non-linear Variance Inflation Factors

Description
A list of data.frames and useful plots for user evaluations of the randomForest hyperparameter mtry.

Usage
robust_vifs(formula, data, model = randomForest, log10 = TRUE, num_var, ...)


Arguments

- **formula**: an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted.
- **data**: an optional data frame containing the variables in the model. By default the variables are taken from the environment which the model is called from.
- **model**: Model to use for extraction partial correlations. Possible model choices are rpart.
- **log10**: Applies a log10 transformation to VIFs when True. Default is True.
- **num_var**: Optional integer argument for reducing the number of variables to the top 'num_var'. Should be an integer between 1 and the total number of predictor variables in the model or it should be a positive proportion of variables desired.
- **...**: Additional arguments to be passed to models as needed.

Value

A list of data.frames and useful plots for user evaluations of VIFs.

Examples

```r
rv <- robust_vifs(Petal.Length ~ ., data = iris[-5], model = lm)
rv
```
Index

* boston
  boston, 2
* lichen
  lichen, 4

boston, 2
formula, 6–8
ggvip, 3
lichen, 4
mtry_compare, 6
partial_cor, 7
robust_vifs, 7