Package 'vtype'

May 14, 2021

Title Estimates the Variable Type in Error Afflicted Data

Version 0.8

Author Andreas Schulz, PhD [aut, cre]

Maintainer Andreas Schulz <ades-s@web.de>

Description Estimates the type of variables in non-quality controlled data. The prediction is based on a random forest model, trained on over 5000 medical variables with accuracy of 99%. The accuracy can hardy depend on type and coding style of data.

Depends R (>= 4.0.0), randomForest

Imports stats

Suggests knitr, rmarkdown

VignetteBuilder knitr

License GPL (>= 3.0)

LazyData true

NeedsCompilation no

Repository CRAN

Date/Publication 2021-05-14 10:30:02 UTC

R topics documented:

vtype-package *Estimates the Variable Type in Error Afflicted Data.*

Description

Estimates the type of variables in non-quality controlled data. The prediction is based on pre-trained random forest model, trained on over 5000 medical variables with OOB accuracy of 9999pct. The accuracy can hardy depend on type and coding style of data.

Details

Package:	vtype
Type:	Package
Version:	0.8
Date:	2021-05-13
License:	GPL version 3 or newer

Author(s)

Andreas Schulz, PhD Maintainer: Andreas Schulz <ades-s@web.de>

sim_nqc_data *Artificial data, that imitates non-quality controlled data*

Description

The data set 'sim_nqc_data' contans 100 observations and 14 variables with some not well formatted and missing values. The data is complete artificial and only intended as an application example for the package.

Usage

```
sim_nqc_data
```

Format

A data frame with 100 observations on 11 (character) variables.

Author(s)

Andreas Schulz

Examples

```
head(sim_nqc_data)
```

---

vtype *Estimates the Variable Type in Error Afflicted Data.*

Description

Estimates the type of variables in not quality controlled data.

Usage

```
vtype(data, qvalue=0.75, miss_values=NULL)
```

Arguments

data	a data frame.
qvalue	Quality value from 0.1 to 1, specifies the proportion of data assumed to be well formatted. The default value of 0.75 works very well most of the time. If the quality of the data is very poor, the q-value can be reduced. If the sample size is very small, it can be increased to use a greater portion of data.
miss_values	a character vector of values considered to be invalid (missing). Important, if missing values were coded as -9 or 9999, otherwise it looks like valid numeric values. Values as NA, NaN, Inf, -Inf, NULL and spaces are automatic considered as invalid (missing) values.

Details

The prediction is based on a pre-trained random forest model, trained on over 5000 medical variables with OOB accuracy of 99pct. The accuracy depends heavily on the type and coding style of data. For example, often categorical variables are coded as integers 1 to x, if the number of categories is very large, there is no way to distinguish it from a continuous integer variable. Some types are per definition very sensitive to errors in data, like ID, missing or constant, where a single alternative non-missing value makes it not constant or not missing anymore. The data is assumed to be cross sectional, where ID is unique (no multiple entries per ID).

Value

A data frame with following entries

- variable: name of the variable
- type: estimated variable type
- probability: probability for estimated type
- format: format of the variable (depending on the type)

- class: broader categorization of type
- alternative: posible alternative type with lesser probability
- n: number of non-missing values
- missings: number of missing values

Examples

```
# Application to a sample data set included in the package.

vtype(sim_nqc_data, miss_values='9999')
```

Index