Package 'rattle.data'

June 26, 2017

Title Rattle Datasets

Version 1.0.2

Date 2017-06-26

Author Graham Williams

Maintainer Graham Williams <Graham.Williams@togaware.com>

Description Contains the datasets used as default examples
by the rattle package. The datasets themselves can be used
independently of the rattle package to illustrate analytics,
data mining, and data science tasks.

Depends R (>= 2.10)

Suggests rattle

LazyData yes

License GPL-3

NeedsCompilation no

Repository CRAN

Date/Publication 2017-06-26 06:12:10 UTC

R topics documented:

audit 2
weather 3
weatherAUS 5
wine 7

Index **9**

audit *Sample dataset to illustrate Rattle functionality.*

Description

The audit dataset is an artificially constructed dataset that has some of the characteristics of a true financial audit dataset for modelling productive and non-productive audits of a person's financial statement. A productive audit is one which identifies errors or inaccuracies in the information provided by a client. A non-productive audit is usually an audit which found all supplied information to be in order.

The audit dataset is used to illustrate binary classification. The target variable is identified as `TARGET_Adjusted`.

The dataset is quite small, consisting of just 2000 entities. Its primary purpose is to illustrate modelling in Rattle, so a minimally sized dataset is suitable.

The dataset itself is derived from publicly available data (which has nothing to do with audits).

Format

A data frame. In line with data mining terminology we refer to the rows of the data frame (or the observations) as entities. The columns are refered to as variables. The entities represent people in this case. We describe the variables here:

`ID` This is a unique identifier for each person.

`Age` The age.

`Employment` The type of employment.

`Education` The highest level of education.

`Marital` Current marital status.

`Occupation` The type of occupation.

`Income` The amount of income declared.

`Gender` The persons gender.

`Deductions` Total amount of expenses that a person claims in their financial statement.

`Hours` The average hours worked on a weekly basis.

`IGNORE_Accounts` The main country in which the person has most of their money banked. Note that the variable name is prefixed with IGNORE. This is recognised by Rattle as the default role for this variable.

`RISK_Adjustment` This variable records the monetary amount of any adjustment to the person's financial claims as a result of a productive audit. This variable, which should not be treated as an input variable, is thus a measure of the size of the risk associated with the person.

`TARGET_Adjusted` The target variable for modelling (generally for classification modelling). This is a numeric field of class integer, but limited to 0 and 1, indicating non-productive and productive audits, respectively. Productive audits are those that result in an adjustment being made to a client's financial statement.

weather *Sample dataset of daily weather observations from Canberra airport in Australia.*

Description

One year of daily weather observations collected from the Canberra airport in Australia was obtained from the Australian Commonwealth Bureau of Meteorology and processed to create this sample dataset for illustrating data mining using R and Rattle.

The data has been processed to provide a target variable `RainTomorrow` (whether there is rain on the following day - No/Yes) and a risk variable `RISK_MM` (how much rain recorded in millimetres). Various transformations were performed on the source data. The dataset is quite small and is useful only for repeatable demonstration of various data science operations.

Usage

```
weather
```

Format

The `weather` dataset is a data frame containing one year of daily observations from a single weather station (Canberra).

`Date` The date of observation (a Date object).

`Location` The common name of the location of the weather station.

`MinTemp` The minimum temperature in degrees celsius.

`MaxTemp` The maximum temperature in degrees celsius.

`Rainfall` The amount of rainfall recorded for the day in mm.

`Evaporation` The so-called Class A pan evaporation (mm) in the 24 hours to 9am.

`Sunshine` The number of hours of bright sunshine in the day.

`WindGustDir` The direction of the strongest wind gust in the 24 hours to midnight.

`WindGustSpeed` The speed (km/h) of the strongest wind gust in the 24 hours to midnight.

`Temp9am` Temperature (degrees C) at 9am.

`RelHumid9am` Relative humidity (percent) at 9am.

`Cloud9am` Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.

`WindSpeed9am` Wind speed (km/hr) averaged over 10 minutes prior to 9am.

`Pressure9am` Atmospheric pressure (hpa) reduced to mean sea level at 9am.

`Temp3pm` Temperature (degrees C) at 3pm.

`RelHumid3pm` Relative humidity (percent) at 3pm.

`Cloud3pm` Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values.

`WindSpeed3pm` Wind speed (km/hr) averaged over 10 minutes prior to 3pm.

`Pressure3pm` Atmospheric pressure (hpa) reduced to mean sea level at 3pm.

`ChangeTemp` Change in temperature.

`ChangeTempDir` Direction of change in temperature.

`ChangeTempMag` Magnitude of change in temperature.

`ChangeWindDirect` Direction of wind change.

`MaxWindPeriod` Period of maximum wind.

`RainToday` Integer: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0.

`TempRange` Difference between minimum and maximum temperatures (degrees C) in the 24 hours to 9am.

`PressureChange` Change in pressure.

`RISK_MM` The amount of rain. A kind of measure of the "risk".

`RainTomorrow` The target variable. Did it rain tomorrow?

Author(s)

<Graham.Williams@togaware.com>

Source

The daily observations are available from http://www.bom.gov.au/climate/data. Copyright Commonwealth of Australia 2010, Bureau of Meteorology.

Definitions adapted from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

References

Package home page: http://rattle.togaware.com. Data source: http://www.bom.gov.au/climate/dwo/ and http://www.bom.gov.au/climate/data.

See Also

weatherAUS, audit.

weatherAUS *Daily weather observations from multiple Australian weather stations.*

Description

Daily weather observations from multiple locations around Australia, obtained from the Australian Commonwealth Bureau of Meteorology and processed to create this realtively large sample dataset for illustrating analytics, data mining, and data science using R and Rattle.

The data has been processed to provide a target variable `RainTomorrow` (whether there is rain on the following day - No/Yes) and a risk variable `RISK_MM` (how much rain recorded in millimeters). Various transformations are performed on the data.

The `weatherAUS` dataset is regularly updated an updates of this package usually correspond to updates to this dataset. The data is updated from the Bureau of Meteorology web site.

The `locationsAUS` dataset records the location of each weather station.

The source dataset is Copyright by the Australian Commonwealth Bureau of Meteorology and is used with permission.

A CSV version of this dataset is available as `https://rattle.togaware.com/weatherAUS.csv`.

Usage

```
weatherAUS
```

Format

The `weatherAUS` dataset is a data frame containing over 140,000 daily observations from over 45 Australian weather stations.

`Date` The date of observation (a Date object).

`Location` The common name of the location of the weather station.

`MinTemp` The minimum temperature in degrees celsius.

`MaxTemp` The maximum temperature in degrees celsius.

`Rainfall` The amount of rainfall recorded for the day in mm.

`Evaporation` The so-called Class A pan evaporation (mm) in the 24 hours to 9am.

`Sunshine` The number of hours of bright sunshine in the day.

`WindGustDir` The direction of the strongest wind gust in the 24 hours to midnight.

`WindGustSpeed` The speed (km/h) of the strongest wind gust in the 24 hours to midnight.

`Temp9am` Temperature (degrees C) at 9am.

`RelHumid9am` Relative humidity (percent) at 9am.

`Cloud9am` Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.

`WindSpeed9am` Wind speed (km/hr) averaged over 10 minutes prior to 9am.

`Pressure9am` Atmospheric pressure (hpa) reduced to mean sea level at 9am.

`Temp3pm` Temperature (degrees C) at 3pm.

`RelHumid3pm` Relative humidity (percent) at 3pm.

`Cloud3pm` Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values.

`WindSpeed3pm` Wind speed (km/hr) averaged over 10 minutes prior to 3pm.

`Pressure3pm` Atmospheric pressure (hpa) reduced to mean sea level at 3pm.

`ChangeTemp` Change in temperature.

`ChangeTempDir` Direction of change in temperature.

`ChangeTempMag` Magnitude of change in temperature.

`ChangeWindDirect` Direction of wind change.

`MaxWindPeriod` Period of maximum wind.

`RainToday` Integer: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0.

`TempRange` Difference between minimum and maximum temperatures (degrees C) in the 24 hours to 9am.

`PressureChange` Change in pressure.

`RISK_MM` The amount of rain. A kind of measure of the "risk".

`RainTomorrow` The target variable. Did it rain tomorrow?

Author(s)

<Graham.Williams@togaware.com>

Source

Observations were drawn from numerous weather stations. The daily observations are available from http://www.bom.gov.au/climate/data.

Definitions adapted from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

References

Package home page: http://rattle.togaware.com. Data source: http://www.bom.gov.au/climate/dwo/ and http://www.bom.gov.au/climate/data.

See Also

weather, audit.

wine *The wine dataset from the UCI Machine Learning Repository.*

Description

The wine dataset contains the results of a chemical analysis of wines grown in a specific area of Italy. Three types of wine are represented in the 178 samples, with the results of 13 chemical analyses recorded for each sample. The Type variable has been transformed into a categoric variable.

The data contains no missing values and consits of only numeric data, with a three class target variable (Type) for classification.

Usage

```
wine
```

Format

A data frame containing 178 observations of 13 variables.

`Type` The type of wine, into one of three classes, 1 (59 obs), 2(71 obs), and 3 (48 obs).

`Alcohol` Alcohol

`Malic` Malic acid

`Ash` Ash

`Alcalinity` Alcalinity of ash

`Magnesium` Magnesium

`Phenols` Total phenols

`Flavanoids` Flavanoids

`Nonflavanoids` Nonflavanoid phenols

`Proanthocyanins` Proanthocyanins

`Color` Color intensity.

`Hue` Hue

`Dilution` D280/OD315 of diluted wines.

`Proline` Proline

Source

The data was downloaded from the UCI Machine Learning Repository.

It was read as a CSV file with no header using `read.csv`. The columns were then given the appropriate names using `colnames` and the Type was transformed into a factor using `as.factor`. The compressed R data file was saved using `save`:

```
UCI <- "http://archive.ics.uci.edu/ml"
REPOS <- "machine-learning-databases"
wine.url <- sprintf("
wine <- read.csv(wine.url, header=FALSE)
colnames(wine) <- c('Type', 'Alcohol', 'Malic', 'Ash',
                    'Alcalinity', 'Magnesium', 'Phenols',
                    'Flavanoids', 'Nonflavanoids',
                    'Proanthocyanins', 'Color', 'Hue',
                    'Dilution', 'Proline')
wine$Type <- as.factor(wine$Type)
save(wine, file="wine.Rdata", compress=TRUE)
```

References

Asuncion, A. & Newman, D.J. (2007). *UCI Machine Learning Repository* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

Index

∗Topic **datasets**
audit, 2
weather, 3
weatherAUS, 5
wine, 7

as.factor, *7*
audit, 2, *4*, *6*

colnames, *7*

locationsAUS *(*weatherAUS*)*, 5

read.csv, *7*

save, *7*

weather, 3, *6*
weatherAUS, *4*, 5
wine, 7