Package 'LogisticEnsembles'

February 26, 2026

Type Package

Title Automatically Runs 18 Logistic Models-14 Individual Logistic Models and 4 Ensembles of Models

Version 1.0.2

Description Automatically returns results from 18 logistic models including 14 individual logistic models and 4 logistic ensembles of models. The package also returns 25 plots, 5 tables, and a summary report. The package automatically builds all 18 models, reports all results, and provides graphics to show how the models performed. This can be used for a wide range of data, such as sports or medical data. The package includes medical data (the Pima Indians data set), and information about the performance of Lebron James. The package can be used to analyze many other examples, such as stock market data. The package automatically returns many values for each model, such as True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate, Positive Predictive Value, Negative Predictive Value, F1 Score, Area Under the Curve. The package also returns 36 Receiver Operating Characteristic (ROC) curves for each of the 18 models.

License MIT + file LICENSE

Depends adabag, arm, brnn, C50, car, caret, corrplot, Cubist, doParallel, dplyr, e1071, gam, gbm, ggplot2, ggplotify, glmnet, graphics, gridExtra, gt, htmltools, htmlwidgets, ipred, klaR, MachineShop, magrittr, MASS, mda, parallel, pls, pROC, purrr, R (>= 4.1.0), randomForest, ranger, reactable, readr, rpart, scales, stats, tidyr, tree, utils, xgboost

Encoding UTF-8

LazyData true

RoxygenNote 7.3.3

Suggests knitr, rmarkdown

VignetteBuilder knitr

URL https://github.com/InfiniteCuriosity/LogisticEnsembles

BugReports https://github.com/InfiniteCuriosity/LogisticEnsembles/issues

NeedsCompilation no

Author Russ Conte [aut, cre, cph]

Maintainer Russ Conte <russconte@mac.com>

Repository CRAN

Date/Publication 2026-02-26 21:50:24 UTC

Contents

`Cervical_cancer` *Cervical_cancer-This data set predicts a patient's risk of cervical cancer based on behavior reports*

Description

"The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients. Several patients decided not to answer some of the questions because of privacy concerns (missing values)." I cleaned up the data so there are no missing data points, nor any NAs.

This data set has 858 observations of 34 variables. The 34th column, 'Biopsy' is the target column.

Age Age

Number.of.sexual.partners Number of reported sexual partners

First.sexual.intercourse Age at first sexual intercourse

Num.of.pregnancies Reported number of pregnancies

Smokes Whether the subject smokes

Smokes..years. The number of years the subject reported smoking

Smokes..packs.year. The number of packs of cigarettes the subject reports smoking each year

Hormonal.Contraceptives If the subject is using hormonal contraceptives

Hormonal.Contraceptives..years. Number of years the subject reports using hormonal contraceptives

IUD Does the subject use an IUD?

IUD..years. Number of years the subject reports using an IUD

STDs Does the patient have STDs?

STDs..number. Number of STDs

STDs.condylomatosis Does the patient have condylomatosis?

STDs.cervical.condylomatosis Does the patient have cervical condylomatosis?

STDs.vaginal.condylomatosis Does the patient have vaginal condylomatosis?

STDs.vulvo.perineal.condylomatosis Does the patient have vulvo perineal condylomatosis?

STDs.syphilis Does the patient have Syphilis?

STDs.pelvic.inflammatory.disease Does the patient have pelvic inflammatory disease?

STDs.genital.herpes Does the patient have genitial herpes?

STDs.molluscum.contagiosum Does the patient have molluscum contagiosum?

AIDS Does the patient have AIDS?

STDs.Hepatitis.B Does the patient have hepatitis B?

STDs..Number.of.diagnosis Number of diagnoses of STDs

Dx.Cancer Does the patient have a diagnosis of cancer?

Dx.CIN Does the patient have a diagnosis of CIN?

Dx.HPV Does the patient have a diagnosis of HPV?

Dx What is the patient's diagnosis?

Hinselmann Hinselmann

Schiller Schiller

Citology Citology

Biopsy The target column, 1 = yes, 0 = no

Usage

```
Cervical_cancer
```

Format

An object of class `data.frame` with 858 rows and 34 columns.

Source

https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors

`Diabetes` *Diabetes—A logistic data set, determining whether a woman tested positive for diabetes. 100 percent accurate results are possible using the logistic function in the Ensembles package.*

Description

"This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset."

This data set is from www.kaggle.com. The original notes on the website state: Context "This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage." Content "The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. Acknowledgements Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261–265). IEEE Computer Society Press.

Pregnancies Number of time pregnant

Glucose Plasma glucose concentration a 2 hours in an oral glucose tolerance test

BloodPressure Diastolic blood pressure (mm Hg)

SkinThickness Triceps skin fold thickness (mm)

Insulin 2-Hour serum insulin (mu U/ml)

BMI Body mass index (weight in kg/(height in m)^2)

DiabetesPedigreeFunction Diabetes pedigree function

Age Age (years)

Outcome Class variable (0 or 1) 268 of 768 are 1, the others are 0

Usage

```
Diabetes
```

Format

An object of class `data.frame` with 768 rows and 9 columns.

Source

<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data>

German_Credit_Risk *German_Credit_Risk-This dataset classifies people described by a set of attributes as good or bad credit risks. #'*

Description

This data set originally came from Professor Hofmann, and is available in several locations, including the UCI Machine Learning Repository I cleaned the data set up, which included naming each of the columns, and removing white spaces from the names of the columns.

The data set has 999 observations of 21 columns of data.The 21st column, "Class" is the target column in the data. Acknowledgements https://dutangc.github.io/CASdatasets/reference/credit.html

Attribute1 Status of existing checking account

Attribute2 Duration (in months)

Attribute3 Credit history

Attribute4 Purpose

Attribute5 Credit amount

Attribute6 Savings accounts/bonds

Attribute7 Present employment since

Attribute8 Installment rate in percentage of disposable income

Attribute9 Personal status and sex

Attribute10 Other debtors / guarantors

Attribute11 Present residence since

Attribute12 Property

Attribute13 Age (in years)

Attribute14 Other installment plans

Attribute15 Housing

Attribute16 Number of existing credits at this bank

Attribute17 Job

Attribute18 Number of people being liable to provide maintenance for

Attribute19 Telephone

Attribute20 Foreign worker

Class 1 = Good, 0 = Bad

Usage

```
German_Credit_Risk
```

Format

An object of class `data.frame` with 999 rows and 21 columns.

Source

https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

`Lebron` *Lebron—A logistic data set, with the result indicating whether or not Lebron scored on each shot in the data set.*

Description

This dataset opens the door to the intricacies of the 2023 NBA season, offering a profound understanding of the art of scoring in professional basketball.

Usage

```
Lebron
```

Format

An object of class `data.frame` with 1533 rows and 12 columns.

Details

top The vertical position on the court where the shot was taken

left The horizontal position on the court where the shot was taken

date The date when the shot was taken. (e.g., Oct 18, 2022)

qtr The quarter in which the shot was attempted, typically represented as "1st Qtr," "2nd Qtr," etc.

time_remaining The time remaining in the quarter when the shot was attempted, typically displayed as minutes and seconds (e.g., 09:26).

result Indicates whether the shot was successful, with "TRUE" for a made shot and "FALSE" for a missed shot

shot_type Describes the type of shot attempted, such as a "2" for a two-point shot or "3" for a three-point shot

distance_ft The distance in feet from the hoop to where the shot was taken

lead Indicates whether the team was leading when the shot was attempted, with "TRUE" for a lead and "FALSE" for no lead

lebron_team_score The team's score (in points) when the shot was taken

opponent_team_score The opposing team's score (in points) when the shot was taken

opponent The abbreviation for the opposing team (e.g., GSW for Golden State Warriors)

team The abbreviation for LeBron James's team (e.g., LAL for Los Angeles Lakers)

season The season in which the shots were taken, indicated as the year (e.g., 2023)

color Represents the color code associated with the shot, which may indicate shot outcomes or other characteristics (e.g., "red" or "green")

@source <https://www.kaggle.com/datasets/dhavalrupapara/nba-2023-player-shot-dataset>

`Logistic` *logistic—function to perform logistic analysis and return the results to the user.*

Description

logistic—function to perform logistic analysis and return the results to the user.

Usage

```
Logistic(
  data,
  colnum,
  numresamples,
  remove_VIF_greater_than,
  remove_data_correlations_greater_than,
  remove_ensemble_correlations_greater_than,
  save_all_trained_models = c("Y", "N"),
  save_all_plots = c("Y", "N"),
  set_seed = c("Y", "N"),
 how_to_handle_strings = c(0("none"), 1("factor levels"), 2("One-hot encoding"),
    3("One-hot encoding with jitter")),
  do_you_have_new_data = c("Y", "N"),
  stratified_column_number,
  use_parallel = c("Y", "N"),
  train_amount,
  test_amount,
  validation_amount
)
```

Arguments

`data`	data can be a CSV file or within an R package, such as MASS::Pima.te
`colnum`	the column number with the logistic data
`numresamples`	the number of resamples
`remove_VIF_greater_than`	Removes features with VIGF value above the given amount (default = 5.00)
`remove_data_correlations_greater_than`	Enter a number to remove correlations in the initial data set (such as 0.98)
`remove_ensemble_correlations_greater_than`	Enter a number to remove correlations in the ensembles
`save_all_trained_models`	"Y" or "N". Places all the trained models in the Environment
`save_all_plots`	Options to save all plots
`set_seed`	Asks the user to set a seed to create reproducible results

`how_to_handle_strings`
: 0: No strings, 1: Factor values

`do_you_have_new_data`
: "Y" or "N". If "Y", then you will be asked for the new data

`stratified_column_number`
: 0 if no stratified random sampling, or column number for stratified random sampling

`use_parallel`
: "Y" or "N" for parallel processing

`train_amount`
: set the amount for the training data

`test_amount`
: set the amount for the testing data

`validation_amount`
: Set the amount for the validation data

Value

a real number

`SAHeart` *SAHeart data*

Description

This is the South African heart disease data originally published in Elements of Statistical Learning, see https://rdrr.io/cran/ElemStatLearn/man/SAheart.html

Usage

```
SAHeart
```

Format

SAHeart

sbp Systolic blood pressure

tobacco cumulative tobacco (kg)

ldl low density lipoprotein cholesterol

adiposity a numeric vector

famhist family history of heart disease, a factor with levels Absent Present

typea type-A behavior

obesity a numeric vector

alcohol current alcohol consumption

age age at onset

chd response, coronary heart disease

Source

Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J. and Ferreira, J. (1983). Coronary risk factor screening in three rural communities, South African Medical Journal 64: 430–436.

Index