Package 'LogisticEnsembles'

April 1, 2025

Type Package

Title Automatically Runs 36 Logistic Models (Individual and Ensembles)

Version 0.5.0

Description Automatically returns 36 logistic models including 23 individual models and 13 ensembles of models of logistic data. The package also returns 10 plots, 5 tables, and a summary report. The package automatically
builds all 36 models, reports all results, and provides graphics to show how the models performed. This can be used for a wide range of data sets. 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 36 models.

License MIT + file LICENSE

Depends adabag, arm, brnn, C50, car, corrplot, Cubist, doParallel,
dplyr, e1071, gam, gbm, ggplot2, ggplotify, graphics,
gridExtra, gt, ipred, klaR, MachineShop, magrittr, MASS, mda,
parallel, pls, pROC, purrr, R (>= 2.10), randomForest, ranger,
reactable, reactablefmtr, readr, rpart, scales, stats, tidyr,
tree, utils, xgboost

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

Suggests knitr, rmarkdown

VignetteBuilder knitr

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

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

NeedsCompilation no

Contents

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>

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_ensemble_correlations_greater_than,
  save_all_trained_models = c("Y", "N"),
  save_all_plots = c("Y", "N"),
  how_to_handle_strings = c("0", "1"),
  do_you_have_new_data = c("Y", "N"),
  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_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
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
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