Package 'NumericEnsembles'

April 10, 2025

Title Automatically Runs 23 Individual and 17 Ensembles of Models

Version 0.7.0

Depends Cubist, Metrics, arm, brnn, broom, car, caret, corrplot, doParallel, dplyr, e1071, earth, gam, gbm, ggplot2, glmnet, graphics, grDevices, gridExtra, ipred, leaps, nnet, parallel, pls, purrr, randomForest, reactable, reactablefmtr, readr, rpart, stats, tidyr, tree, utils, xgboost, R (>= 4.1.0)

Description Automatically runs 23 individual models and 17 ensembles on numeric data. The package automatically returns complete results on all 40 models, 25 charts, multiple tables. The user simply provides the data, and answers a few questions (for example, how many times would you like to resample the data). From there the package randomly splits the data into train, test and validation sets, builds models on the training data, makes predictions on the test and validation sets, measures root mean squared error (RMSE), removes features above a user-set level of Variance Inflation Factor, and has several optional features including scaling all numeric data, four different ways to handle strings in the data. Perhaps the most significant feature is the package's ability to make predictions using the 40 pre trained models on totally new (untrained) data if the user selects that feature. This feature alone represents a very effective solution to the issue of reproducibility of models in data science. The package can also randomly resample the data as many times as the user sets, thus giving more accurate results than a single run. The graphs provide many results that are not typically found. For example, the package automatically calculates the Kolmogorov-Smirnov test for each of the 40 models and plots a bar chart of the results, a bias bar chart of each of the 40 models, as well as several plots for exploratory data analysis (automatic histograms of the numeric data, automatic histograms of the numeric data). The package also automatically creates a summary report that can be both sorted and searched for each of the 40 models, including RMSE, bias, train RMSE, test RMSE, validation RMSE, overfitting and duration. The best results on the holdout data typically beat the best results in data science competitions and published results for the same data set.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.2

LazyData true

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

VignetteBuilder knitr

URL http://www.NumericEnsembles.com,
https://github.com/InfiniteCuriosity/NumericEnsembles

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

NeedsCompilation no

Author Russ Conte [aut, cre, cph]

Maintainer Russ Conte <russconte@mac.com>

Repository CRAN

Date/Publication 2025-04-10 21:10:13 UTC

Contents

`Boston_housing` *Boston_housing data*

Description

This is a modified version of the famous Boston housing data set. This data set includes rows 4:209 and 212:506. The data here is complete except for the data use to make New_Boston. The data first appeared in a paper by David Harrison, Jr. and Daniel L. Rubenfeld, Hedonic housing Prices and the demand for clean air. This was published in March, 1978. Journal of Environmental Economics and Management 5(1):81-102. The descriptions below are quoted from the original paper:

crim Crime rate by town. Original data in 1970 FBI data

zn Proportion of a town's residential land zoned for lots greater than 25,000 square feet

indus Proportional non-retail business per town

chas Captures the amenities of a riverside location and thus should be positive

nox Nitrogen oxygen concentrations in part per hundred million

rm Average number of rooms in owner units

age Proportion of owner units built prior to 1940

dis Weighted distances to five employment centers in the Boston region

rad Index of accessibility to radial highways

tax Full property value tax rate ($/$10,000)

ptratio Pupil-teacher ratio by town school district

black Black proportion of population

lstat Proportion of population that is lower status (proportion of adults without some high school education and proportion of male workers classified as laborers)

medv Median value of owner occupied homes, from the 1970 United States census

Usage

```
Boston_housing
```

Format

An object of class `data.frame` with 501 rows and 14 columns.

Source

https://www.law.berkeley.edu/files/Hedonic.PDF

`Concrete` *Concrete - This is the strength of concrete daa set originally posted on UCI*

Description

Concrete is the most important material in civil engineering. The concrete compressive strength is a highly nonlinear function of age and ingredients.

Usage

```
Concrete
```

Format

Concrete A data frame with 1030 rows and 9 columns:

Cement quantitative – kg in a m3 mixture – Input Variable

Blast_Furnace_Slag quantitative – kg in a m3 mixture – Input Variable

Fly_Ash quantitative – kg in a m3 mixture – Input Variable

Water quantitative – kg in a m3 mixture – Input Variable

Superplasticizer quantitative – kg in a m3 mixture – Input Variable

Coarse_Aggregate quantitative – kg in a m3 mixture – Input Variable

Fine_Aggregate quantitative – kg in a m3 mixture – Input Variable

Age Day (1~365) – Input Variable

Strength quantitative – MPa – Output Variable

Source

https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength

---

`Insurance` *Insurance - The data is from UCI*

Description

This dataset contains detailed information about insurance customers, including their age, sex, body mass index (BMI), number of children, smoking status and region. Having access to such valuable insights allows analysts to get a better view into customer behaviour and the factors that contribute to their insurance charges.

Usage

```
Insurance
```

Format

Insurance A data frame with 1338 rows and 7 columns Credit to Bob Wakefield

Age The age of the customer. (Integer)

Children The number of children the customer has. (Integer)

Smoker Whether or not the customer is a smoker. (Boolean)

Region The region the customer lives in. (String)

Charges The insurance charges for the customer. (Float)

Source

https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender

---

`New_Boston` *NewBoston—These are only the five rows c(1:3, 210:211) from Boston Housing data set. This can be used as new data, and the Boston_housing data set as the original. The numeric function will return predictions on the new data.*

Description

This is the first five rows of the Boston housing data set, which have been removed from the Boston data set included here. It is otherwise identical to the Boston data set.

crim Crime rate by town. Original data in 1970 FBI data

zn Proportion of a town's residential land zoned for lots greater than 25,000 square feet

indus Proportional non-retail business per town

chas Captures the amenities of a riverside location and thus should be positive

nox Nitrogen oxygen concentrations in part per hundred million

rm Average number of rooms in owner units

age Proportion of owner units built prior to 1940

dis Weighted distances to five employment centers in the Boston region

rad Index of accessibility to radial highways

tax Full property value tax rate ($/$10,000)

ptratio Pupil-teacher ratio by town school district

black Black proportion of population

lstat Proportion of population that is lower status (proportion of adults without some high school education and proportion of male workers classified as laborers)

medv Median value of owner occupied homes, from the 1970 United States census

Usage

```
New_Boston
```

Format

An object of class `data.frame` with 5 rows and 14 columns.

Source

https://www.law.berkeley.edu/files/Hedonic.PDF

`Numeric` *Numeric—function to automatically build 23 individual models and 17 ensembles then return the results to the user*

Description

Numeric—function to automatically build 23 individual models and 17 ensembles then return the results to the user

Usage

```
Numeric(
  data,
  colnum,
  numresamples,
  remove_VIF_above = 5,
  remove_ensemble_correlations_greater_than = 0.98,
  scale_all_predictors_in_data = c("Y", "N"),
  data_reduction_method = c(0("none"), 1("BIC exhaustive"), 2("BIC forward"),
    3("BIC backward"), 4("BIC seqrep"), 5("Mallows_cp exhaustive"),
    6("Mallows_cp forward"), 7("Mallows_cp backward"), 8("Mallows_cp, seqrep")),
  ensemble_reduction_method = c(0("none"), 1("BIC exhaustive"), 2("BIC forward"),
    3("BIC backward"), 4("BIC seqrep"), 5("Mallows_cp exhaustive"),
    6("Mallows_cp forward"), 7("Mallows_cp backward"), 8("Mallows_cp, seqrep")),
  how_to_handle_strings = c(0("none"), 1("factor levels"), 2("One-hot encoding"),
    3("One-hot encoding with jitter")),
  predict_on_new_data = c("Y", "N"),
  save_all_trained_models = c("Y", "N"),
  save_all_plots = 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::Boston
`colnum`	a column number in your data
`numresamples`	the number of resamples
`remove_VIF_above`	remove columns with Variable Inflation Factor above value chosen by the user
`remove_ensemble_correlations_greater_than`	maximum value for correlations of the ensemble
`scale_all_predictors_in_data`	"Y" or "N" to scale numeric data
`data_reduction_method`	0(none), BIC (1, 2, 3, 4) or Mallow's_cp (5, 6, 7, 8) for Forward, Backward, Exhaustive and SeqRep
`ensemble_reduction_method`	0(none), BIC (1, 2, 3, 4) or Mallow's_cp (5, 6, 7, 8) for Forward, Backward, Exhaustive and SeqRep
`how_to_handle_strings`	0: No strings, 1: Factor values, 2: One-hot encoding, 3: One-hot encoding AND jitter
`predict_on_new_data`	"Y" or "N". If "Y", then you will be asked for the new data

`save_all_trained_models`
: "Y" or "N". If "Y", then places all the trained models in the Environment

`save_all_plots`
: Saves all plots to the working directory

`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

Index