Package ‘audrex’

August 18, 2021

Type Package
Title Automatic Dynamic Regression using Extreme Gradient Boosting
Version 1.0.1
Author Giancarlo Vercellino
Maintainer Giancarlo Vercellino <giancarlo.vercellino@gmail.com>
Description Dynamic regression for time series using Extreme Gradient Boosting with hyper-parameter tuning via Bayesian Optimization.
License GPL-3
Encoding UTF-8
LazyData true
RoxygenNote 7.1.1
Depends R (>= 3.6)
Imports rBayesianOptimization, xgboost, purrr, abind, ggplot2, readr, stringr, lubridate, narray, fANCOVA, imputeTS, scales, tictoc, modeest, moments
URL https://rpubs.com/giancarlo_vercellino/audrex
NeedsCompilation no
Repository CRAN
Date/Publication 2021-08-18 18:10:02 UTC

R topics documented:

audrex ................................................................. 2
bitcoin_gold_oil ....................................................... 5
climate_anomalies ................................................... 6
covid_in_europe ..................................................... 6

Index 8
Description

Dynamic regression for time series using Extreme Gradient Boosting with hyper-parameter tuning via Bayesian Optimization.

Usage

```
audrex(data, targets, past = NULL, future = NULL, deriv = 0, smoother = F, ci = 0.8, windows = 3, internal_holdout = 0.5, nrounds = 100, patience = 10, booster = "gbtree", max_depth = NULL, eta = NULL, gamma = NULL, min_child_weight = NULL, subsample = NULL, colsample_bytree = NULL, lambda = NULL, alpha = NULL, verbose = FALSE, reg = "squarederror", eval_metric = "rmse", starting_date = NULL, time_unit = NULL, dbreak = NULL, min_set = 30, seed = 42, opt_metric = "mae", n_samp = 10, n_search = 5, acq = "ucb", kappa = 2.576, eps = 0, kernel = list(type = "exponential", power = 2)),
```
minmax = FALSE
)

Arguments

data A data frame with time series on columns and possibly a date column (not mandatory)

targets String. Names of ts features to be jointly analyzed: for each feature a distinct model is built using the others as regressors.
past Positive integer. The past dimension with number of time-steps in the past used for the prediction.
future Positive integer. The future dimension with number of time-steps to be predicted
deriv Positive integer. Number of differentiation operations to perform on the original series. 0 = no change; 1: one diff; 2: two diff, and so on.
smoother Logical. Perform optimal smoothing using standard loess. Default: FALSE

ci Confidence interval. Default: 0.8
windows Positive integer. Number of (expanding) windows for cross-validation. Default: 3.
internal_holdout Positive numeric. Holdout percentage for internal xgb validation. Default: 0.5.
nrounds Positive numeric. Number of round for the extreme boosting machine. Look to xgboost for description. Default: 100.
patience Positive integer. Waiting rounds without improvement before xgboost stops. Default: 10
booster String. Optimization methods available are: "gbtree", "gblinear". Default: "gbtree".
max_depth Positive integer. Look to xgboost documentation for description. A vector with one or two positive integer for the search boundaries. The default value (NULL) sets automatically the values in c(1, 10).
eta Positive numeric. Look to xgboost documentation for description. A vector with one or two positive numeric between (0, 1] for the search boundaries. The default value (NULL) sets automatically the values in c(0.001, 1).
gamma Positive numeric. Look to xgboost documentation for description. A vector with one or two positive numeric for the search boundaries. The default value (NULL) sets automatically the values in c(0.001, 100).
min_child_weight Positive numeric. Look to xgboost documentation for description. A vector with one or two positive numeric for the search boundaries. The default value (NULL) sets automatically the values in c(1, 100).
subsample Positive numeric. Look to xgboost documentation for description. A vector with one or two positive numeric between (0, 1] for the search boundaries. The default value (NULL) sets automatically the values in c(0.1, 1).
colsample_bytree
Positive numeric. Look to xgboost documentation for description. A vector with one or two positive numeric between (0, 1] for the search boundaries. The default value (NULL) sets automatically the values in c(0.1, 1).

lambda
Positive numeric. Look to xgboost documentation for description. A vector with one or two positive numeric for the search boundaries. The default value (NULL) sets automatically the values in c(0.1, 100).

alpha
Positive numeric. Look to xgboost documentation for description. A vector with one or two positive numeric for the search boundaries. The default value (NULL) sets automatically the values in c(0.1, 100).

verbose
Logical. Default: TRUE

reg
String. Learning objective function. Options are: "squarederror", "pseudohubererror". Default: "squarederror".

eval_metric
String. Evaluation metric for the boosting algorithm. Options are: "rmse", "mae", "mape". Default: "mae".

starting_date
Date. Initial date to assign temporal values to the series. Default: NULL (progressive numbers).

time_unit
String. Time step of the features, in liberal form: i.e., "20 seconds", "10 week", "1 day". Default: NULL.

dbreak
String. Minimum time marker for the plot x-axis, in liberal form: i.e., "3 months", "1 week", "20 days". Default: NULL.

min_set

seed
Random seed. Default: 42.

opt_metric
String. Parameter for selecting the best model, averaging one-step error across all ts features. Default: "mae".

n_samp

n_search
Positive integer. Number of search steps for the Bayesian Optimization. Default: 5.

acq
String. Parameter for Bayesian Optimization. For reference see rBayesianOptimization documentation. Default: "ucb".

kappa
Positive numeric. Parameter for Bayesian Optimization. For reference see rBayesianOptimization documentation. Default: 2.576.

eps
Positive numeric. Parameter for Bayesian Optimization. For reference see rBayesianOptimization documentation. Default: 0.

kernel
List. Parameter for Bayesian Optimization. For reference see rBayesianOptimization documentation. Default: list(type = "exponential", power = 2).

minmax
Logical. Boolean flag to apply minmax normalization. Default: FALSE.
Value

This function returns a list including:

- **best_par**: the parameter of the best model selected through Bayesian Optimization
- **history**: a table with the sampled models \((n_{\text{samp}} + n_{\text{search}})\), their parameters and optimization metric
- **best_model**: results for the best selected model, including:
  - **errors**: training and testing errors for one-step and sequence for each ts feature \((\text{rmse, mae, mdae, mpe, mape, smape})\)
  - **predictions**: min, max, q25, q50, q75, quantiles at selected ci, mean, sd for each ts feature
  - **pred_stats**: for each predicted time feature, IQR to range, Kullback-Leibler Divergence (compared to previous point in time), upside probability (compared to previous point in time), both averaged across all points in time and compared between the terminal and the first point in the prediction sequence.
- **time_log**

Author(s)

Giancarlo Vercellino <giancarlo.vercellino@gmail.com>

See Also

Useful links:

- [https://rpubs.com/giancarlo_vercellino/audrex](https://rpubs.com/giancarlo_vercellino/audrex)

Examples

```r
audrex(covid_in_europe, "daily_cases", past = 10, future = 5, deriv = 1, n_samp = 5, n_search = 3)
```

---

**bitcoin_gold_oil**  
**bitcoin_gold_oil data set**

Description

A data frame with different time series (prices and volumes) for bitcoin, gold and oil.

Usage

```r
bitcoin_gold_oil

bitcoin_gold_oil
```
covid_in_europe

Format
A data frame with 18 columns and 1827 rows.
A data frame with 18 columns and 1827 rows.

Source
Yahoo Finance
Yahoo Finance

climate_anomalies  climate_anomalies data set

description
A data frame with different two time series on global mean temperature anomalies (GMTA) and
global mean sea level (GMTA).

Usage
climate_anomalies

Format
A data frame with 2 columns and 266 rows.

Source
Datahub.io, Climate-change collection

covid_in_europe  covid_in_europe data set

description
A data frame with daily and cumulative cases of Covid infections and deaths in Europe since
March 2021.
A data frame with daily and cumulative cases of Covid infections and deaths in Europe since
March 2021.

Usage
covid_in_europe
covid_in_europe
Format

A data frame with 5 columns and 163 rows.
A data frame with 5 columns and 163 rows.

Source

www.ecdc.europa.eu
www.ecdc.europa.eu
Index

* datasets
  - bitcoin_gold_oil, 5
  - climate_anomalies, 6
  - covid_in_europe, 6

audrex, 2
audrex-package (audrex), 2

bitcoin_gold_oil, 5
climate_anomalies, 6
covid_in_europe, 6