tsdataleaks: An R Package to Detect Potential Data Leaks in Forecasting Competitions

26 December 2023

Summary

Forecasting competitions are of increasing importance as a mean to learn best practices and gain knowledge. Data leakage is one of the most common issues that can often be found in competitions. Data leaks can happen when the training data contains information about the test data. There are a variety of different ways that data leaks can occur with time series data. For example: i) randomly chosen blocks of time series are concatenated to form a new time series, ii) scale-shifts, iii) repeating patterns in time series, iv) white noise is added in the original time series to form a new time series, etc. This work introduces a novel tool to detect these data leaks. The tsdataleaks package provides simple and computationally efficient algorithm to exploit data leaks in time series data. This paper demonstrates the package design and its power to detect data leakages using recent forecasting competitions data.

Statement of Need

Time series forecasting competitions have played a significant role in the advancement of forecasting practices. Typically, in forecasting competitions, a collection of time series is given to the competitors, and then the competitors submit the forecasts for the required test period of each time series. During the competition period only the training set of each time series is given to the public, and the test set is kept private from the public. Finally, competition organizers evaluate the forecast accuracy comparing the test set of each series and submitted forecasts from the different competitiors. This helps to identify new forecasting techiniques.

Data leakage occur when the training period of the time series includes test period data before officially release the test period of the time series. This idea is illustrated in Figure 1. A and B are two time series. The latter segment of the training set and the subsequent test set within the (B) series is derived from a training segment inherent to series (A). This type of data leak could occur when a randomly chosen blocks of time series are concatenated to form a new time series.

Competitions with data leaks will not be able to reach the original purpose of their competitions. By exploiting data leakage competitors can obtain a top rank in the leader board. Such models look highly accurate within the competition environment but becomes inaccurate when applying the to a data set outside the competition environment. There is an increasing need to examine the potential data leaks in time series before the release of data to public. The tsdataleaks package is designed to identify data leaks in time series.

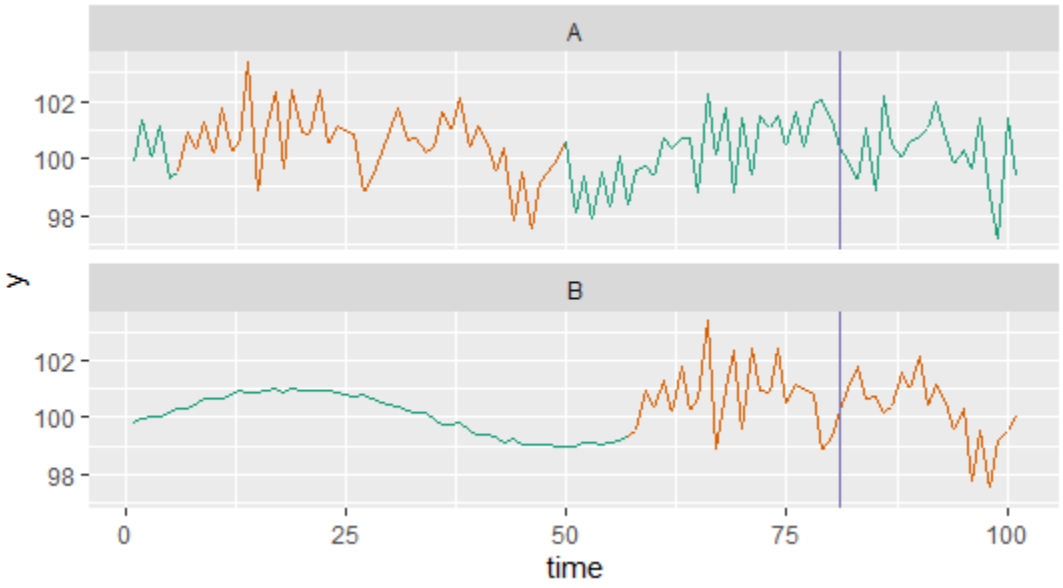

Figure 1: An example of a time series data leak. (A) and (B) are two time series. The purple verticle line separates the training and test parts of the series. The latter segment of the training set and test set of the (B) series comes from a training segment of series (A).

State of the Field in R

Features

The Sections *** demonstrate the utility of these packages as well as walk two examples.

Reproducibility