Out-of-sample Testing Overview

Out-of-sample Testing · https://dev.forecastpro.com/documentation/forecast-pro-trac/out-of-sample-testing/out-of-sample-testing-overview/

A good deal of the empirical knowledge about forecasting has come from comparisons of different methodologies. The M-Competition (Makridakis et al. [1982]), M-3 Competition (Makridakis and Hibon [2000]), M-4 Competition (Makridakis et al. [2019]) and M-5 Competition (Makridakis et al. [2022]) are the largest and most famous of these comparisons. Forecast Pro participated in the M-3 and M-4 competitions and outperformed all other commercial software entrants in both.

The rather simple comparison methodology for the original M-Competition was as follows.

The researchers assembled a collection of 1001 time series of yearly, quarterly and monthly data. The data were obtained from microeconomic, industry-level, macroeconomic and demographic sources. Twenty forecasting methods were tested for the entire sample of 1001 time series and three on a subset of only 111 time series.

A sample of time points (6 for annual series, 8 for quarterly, 18 for monthly) was held out from the end of each time series. Each forecast model was fitted to the remaining data and used to forecast the values of the holdout sample. The forecasts were then compared to the withheld data, and errors computed for each horizon, each time series and each forecast method. The errors were then summarized and analyzed in a variety of ways.

The most significant weakness in this methodology is that it uses only one forecast base for each time series, the last point in the fitting sample. One obtains only a “snapshot” of performance from one point in time. A forecast base just before or after a dramatic event in the data may completely change the results. Furthermore, you obtain only one forecast error for each horizon time from 1 to the end of the fit set. This procedure is referred to as a *static* evaluation.

Forecast Pro implements both a static and a *rolling-base* evaluation. The rolling-base procedure begins in the same way. However, after the forecasts have been made, the model is rolled forward by one period. Forecasts are then made from the new base to the end of the withheld data. This process is repeated until the withheld data sample is exhausted. If 6 data points have been withheld, then you

obtain 6 1-step forecasts, 5 2-step forecasts, 4 3-step forecasts, etc.

The model coefficients are *not* reestimated as each additional data point is assimilated. The forecast model is based entirely upon the original fit set.