Introduction

William K Davis III

2022-06-27

Evaluation

For the purposes of discussing model error rate, let y_t be the observed bike count in period t, $\hat{y}_t$ be the forecast bike count in period t, and $\hat{e}_t = y_t - \hat{y}_t$ be the forecast error at time t.

Model Error Rate

Given our stated use for these models is forecasting (prediction), when discussing model evaluation we must first define the notion of forecast (prediction) error rate. There are a number of different metrics to use for forecast error, each with their own benefits and drawbacks (Hyndman and Koehler [2006]). We will use Mean Absolute Error (MAE) to select the best model from among the list of candidate models. Unlike percent errors, which have the general form $100 \times \frac{\hat{e}}{y_t}$, MAE is defined when $y_t = 0$. Further, squared errors such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are sensitive to outliers. Finally, MAE is on the same scale as the original dataset (number of bikes per hour), which gives it nice properties of interpretability. Therefore, MAE is the best error metric for selecting the best model and estimating test error. For our purposes, we will use the following equation:

$$\text{Mean Absolute Error} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} |e_{ij}|$$

Where $m = 24$ is the number of periods for which bike count is forecast and $n = 50$ is the number of iterations in the cross-validation.

Model Selection and Test Error

Models will be selected and evaluated in the context of their use for forecasting rather than statistical inference. As such, we will use two rounds of time series cross-validation (TSCV), one round for hyperparameter selection within each modeling methodology and one round for selecting the best model from among the different methodologies. TSCV consists of selecting a point or series of points from the dataset as test sets, then selecting all prior points as the training set (Hyndman and Athanasopoulos [2021]). In the example in Figure F, the first iteration trains on the first 5 observations (blue) and generates forecasts on the next 3 observations (green). In the second iteration the model trains on the first 5+3=8 observations and forecasts on the next 3 observations. This continues until the final iteration, where the model trains on all but the last 3 observations and then generates forecasts for the final 3 observations. In this example 5 is the initialization value (the number of training observations in the first iteration), 3 is the step size (the number of observations that are added to the training set each iteration), and 3 is also the horizon (the number of periods for which we generate a forecast). Each of the forecasts is then compared with actual (observed) values to evaluate the forecast.

For this evaluation there will be two stages of TSCV. The first stage will be to select optimal hyperparameters within each modeling method. The second stage will be to select the best model from among the different modeling methods. The first stage will have 50 iterations with a forecast horizon of 27 hours and a step size of 24 hours. The second stage will have 10 iterations with the same horizon and step size. The step size of 24

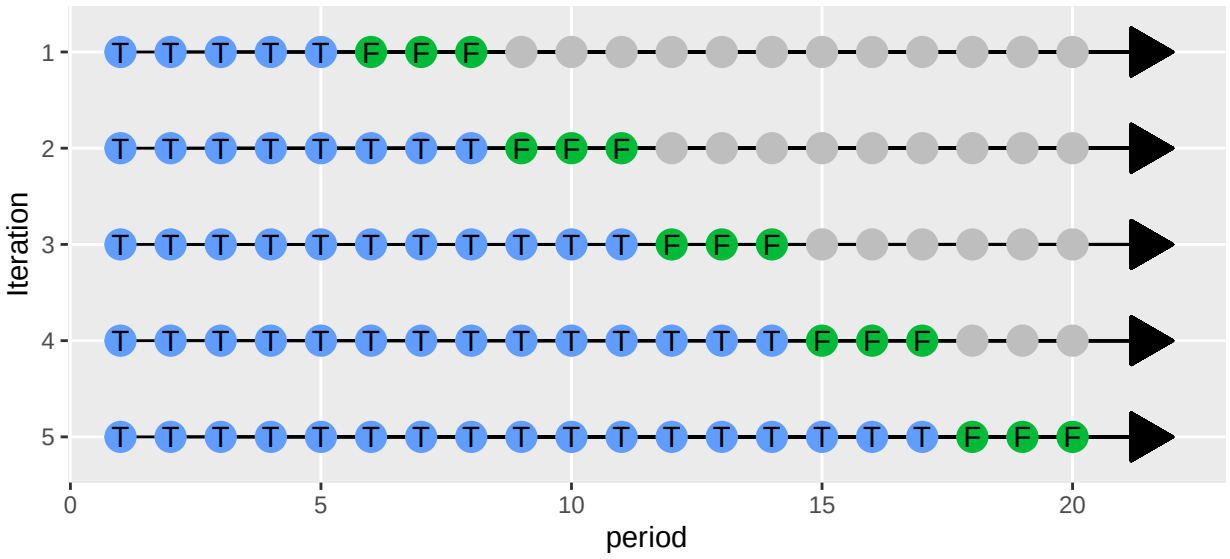

Figure 1: Time series cross validation

hours, aside from having meaning as the length of a day, is meant to avoid possible serial correlation in the errors that can occur with smaller step sizes, such as 1 hour (**?**).

The last training period will be at 8:00pm with forecast periods for 9:00pm to 11:00pm the following day. This serves to simulate real-world usage, where the operator of a bike share program would generate forecasts each day at 8:00pm, after peak demand, for the following day, so that the overnight hours can be used to reallocating bikes based on the forecast for the following day's demand. The first forecast period of the first iteration of the first cross validation stage will be on XX, and the last forecast observation of the final (50th) iteration of the first stage will be on YY. The first forecast observation of the first iteration of the second stage will be on XX and the final forecast observation of the final (10th) iteration of the second stage will be on YY.

References

Rob Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3rd edition, 2021. URL OTexts.com/fpp3.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2006.03.001. URL https://www.sciencedirect.com/science/article/pii/S0169207006000239.