Model Evaluation
Point and Density Forecasts

Romain Lafarguette, Ph.D.   Amine Raboun, Ph.D.

Quants & IMF External Experts

romainlafarguette.github.io/   amineraboun.github.io/

Singapore Training Institute, 19 April 2023

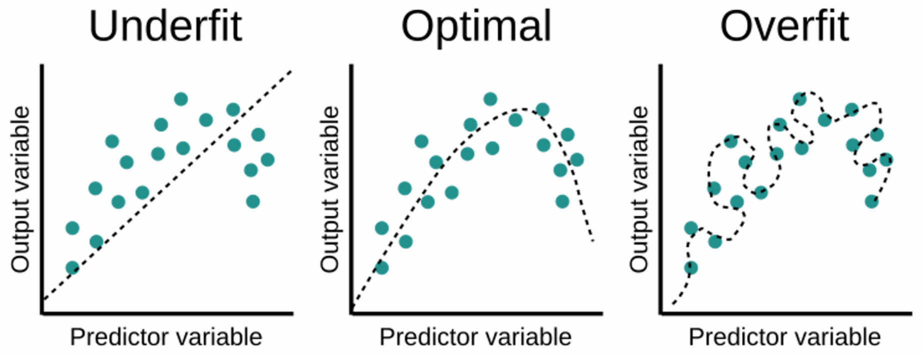
Fitting and Forecasting

Be careful

A model that fits the data well (in sample) might not necessarily forecast well

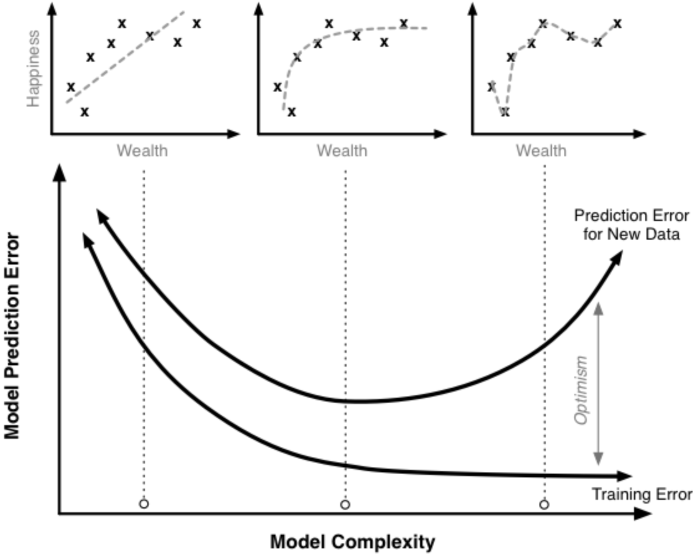
- A perfect in-sample fit can always be obtained by using a model with with enough parameters
- Over-fitting a model to data is just as bad as failing to identify a systematic pattern in the data
- Need to split the model between
- The test set must no be used to *any* aspect of model development or calculation of forecasts
- Forecast accuracy is only based on the test set

Underfit, Optimal, Overfit: Intuition



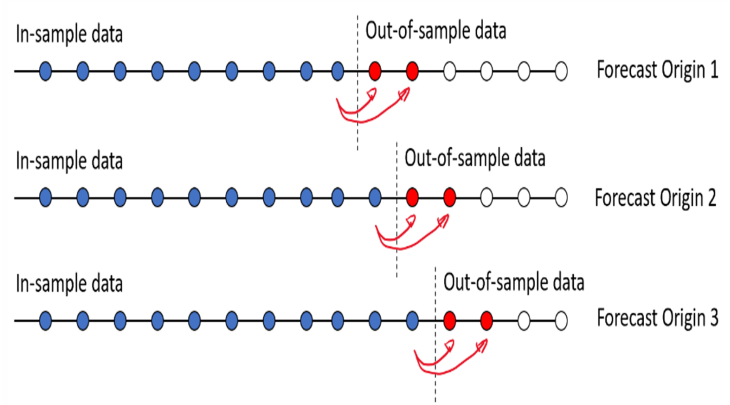
Source: *towardsdatascience*

Underfit, Optimal, Overfit and Model Complexity



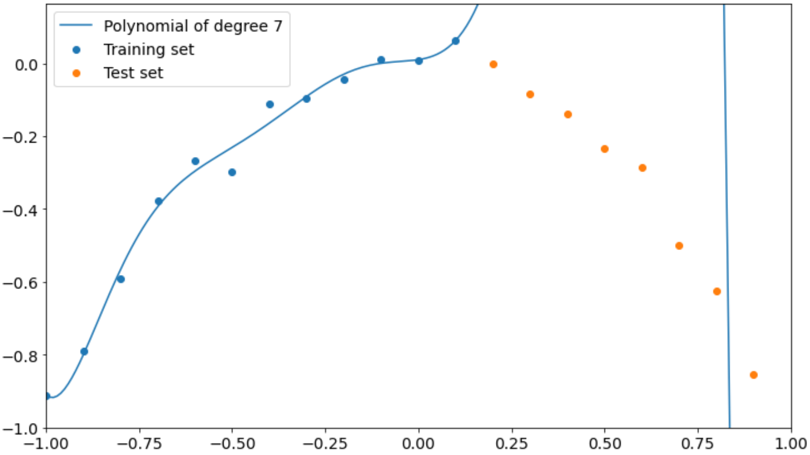
Source: *Scott Fortmann-Roe*

Out of Sample Concept



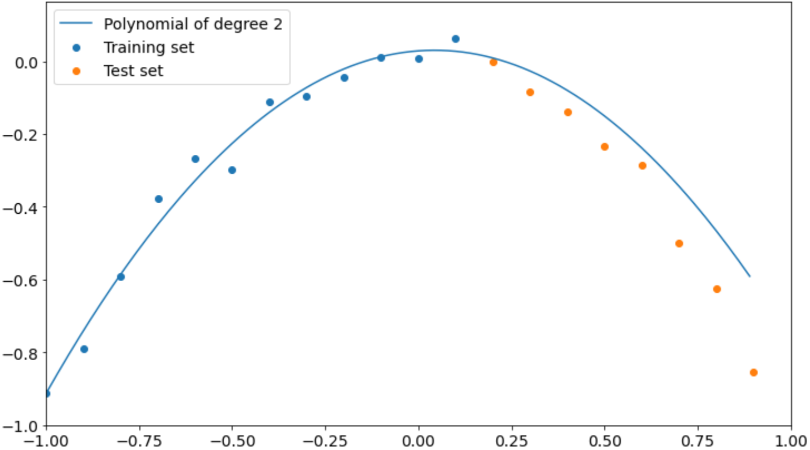
Source: *Author*

Out of Sample Example: Overfit



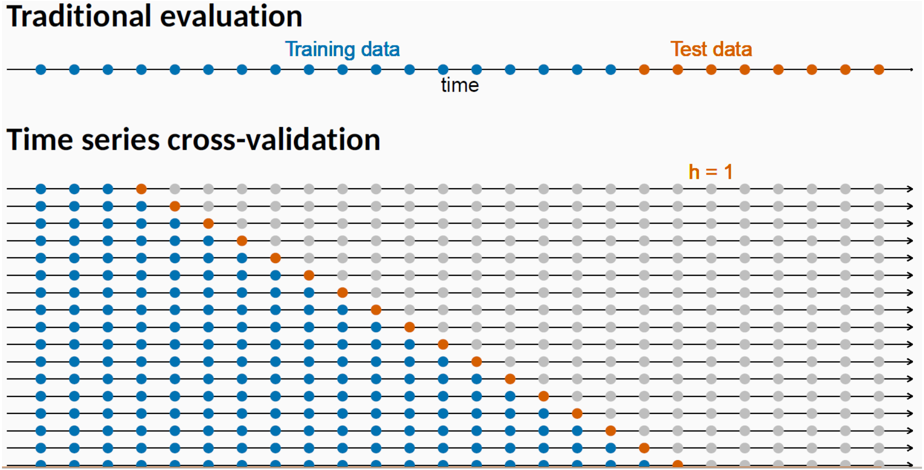
Source: *towardsdatascience.com/an-example-of-overfitting-and-how-to-avoid-it*

Out of Sample Example: Correct Fit

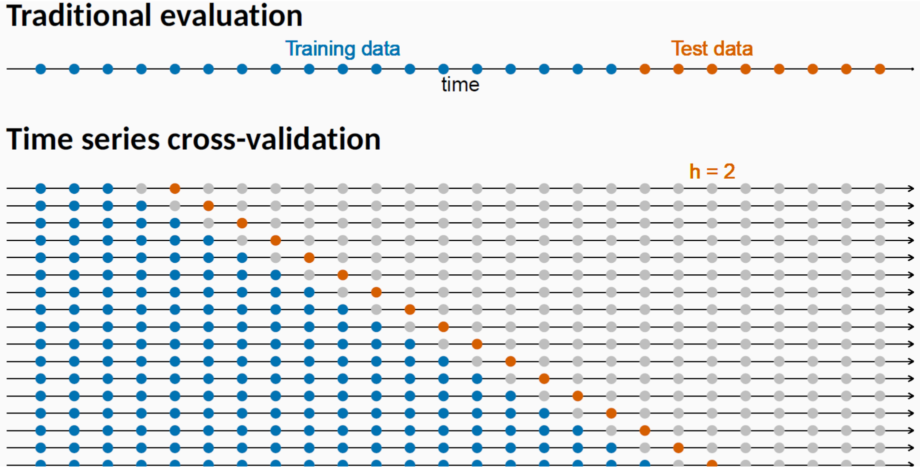


Source: *towardsdatascience.com/an-example-of-overfitting-and-how-to-avoid-it*

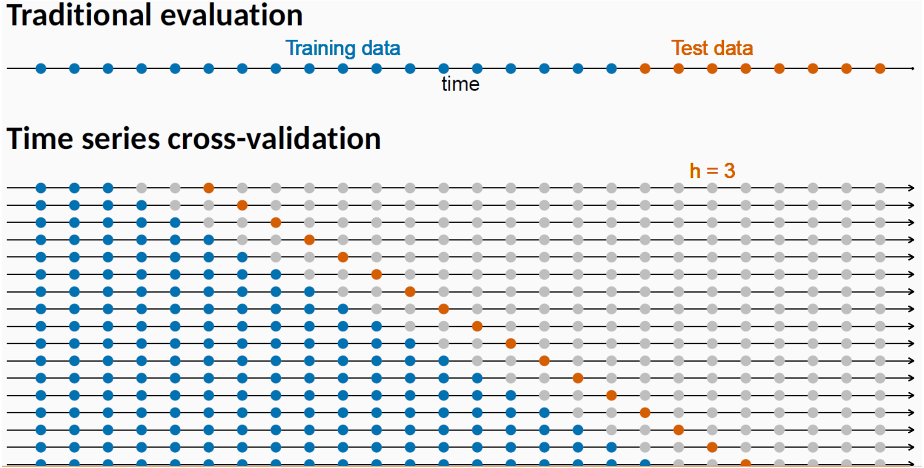
Time Series Cross-Validation



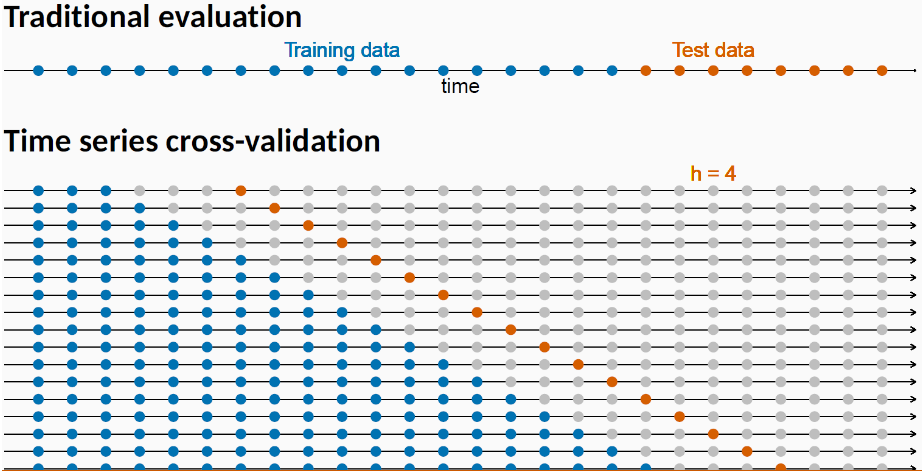
Time Series Cross-Validation



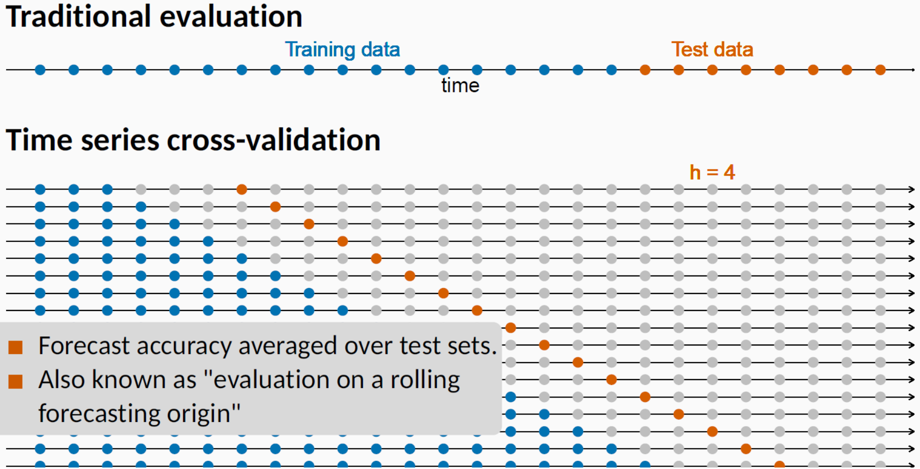
Time Series Cross-Validation



Time Series Cross-Validation



Time Series Cross-Validation



Forecast Errors

- Evaluating point forecasts are relatively straightforward
- Ex-post (after the realization happened), we observe:
  - The true value y_{T+h} that has been realized
  - The expected value $\hat{y_{T+h}}$ that has been generated before, in time t

Definition: Forecast Errors

A forecast error is the ex-post difference between an observed value and its forecast

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h} | Y_T, \ldots, Y_1$$

- Forecast evaluation metrics represent different variations on how to summarize the e_{T+h}
  - Are the forecast errors small on average?
  - Have we observed infrequent but large forecast errors (outliers)?
  - Are the forecast errors evenly distributed across the distribution of y? etc.

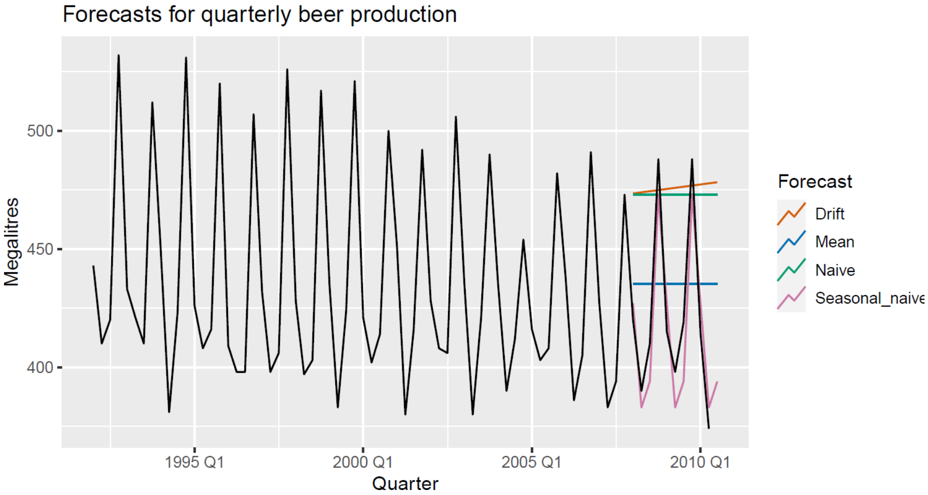
Forecast Errors with Train/Test Sets

Out of Sample

Measuring **accuracy** should be done out of sample. In-sample metrics inform on the how well the model **fits** the data

- The conditional set $Y_T, \ldots, Y_1$ should only be taken from the training dataset
- The true value y_{T+h} is taken from the test set
- Unlike residuals, forecast errors on the test involve multi-step forecasts
- These are the **true** forecast error, as the test data is not used to compute $\hat{y}_{T+h}$

Example: Forecasting Beer Production



Measures of Forecast Accuracy

Main Metrics

- **MAE**: mean absolute errors $\frac{1}{S}\sum_{s\in S}|e_{s,T+h}|$
- **MSE**: mean squared errors $\frac{1}{S}\sum_{s\in S}(e_{s,T+h})^2$
- **MAPE**: mean absolute percentage errors $\frac{1}{S}100 * \sum_{s\in S}\frac{|e_{s,T+h}|}{|y_{s,t+h}|}$
- **RMSE**: root mean squared errors: $\sqrt{\frac{1}{S}\sum_{s\in S}(e_{s,T+h})^2}$

With:

- y_{T+h}: T+h observation, h being the horizon (h = 1, 2, ..., H)
- $\hat{y}_{T+h|T}$: the forecast based on data up to time T
- $e_{T+h} = y_{T+h} - \hat{y}_{T+h|T}$: The forecast errors
- S is the testing sample

Scaling

- MAE, MSE and RMSE are all **scale dependent**
- MAPE is scale independent but is only sensible if $y_t >> 0 \qquad \forall\ t$
- **Most commonly used: Time Cross-Validation with the lowest RMSE**

Nemenyi Test

- We can rank the model by RMSE (or another metric), but are the RMSE significantly different?

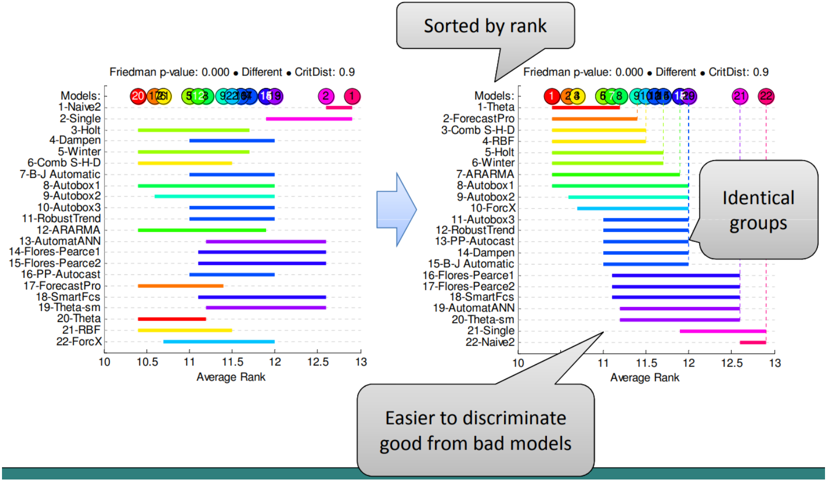
- Maybe Model 1 can have a lower RMSE than Model 2, but the difference in RMSE is non-significant

- In which case, we could pool the two models together

- Use a non-parametric test to test the hypothesis of equal RMSE, with the test statistic:

$$r_{\alpha,K,N} \approx \frac{q_{\alpha,K}}{\sqrt{2}}\sqrt{\frac{K(K+1)}{6N}}$$

Nemenyi Test in Practice



Source: *Nikolaos Kourentzes*

Challenges

- At the difference of point forecasts, density forecasts are never observed
  - We only observe **one** realization of the density
- Hence, for evaluating the quality of the density forecasts, we need to use specific tools
- The **model specification**: is my model "neutral", not over-optimistic, not over-pessimistic?
  - Use a **Probability Integral Transform (PIT) test**
- The **model performance**: attributing high *ex-ante* performance to *ex-post* realizations
  - Use **logscores** and asymmetric logscores

Probability Integral Transform Test (PIT)

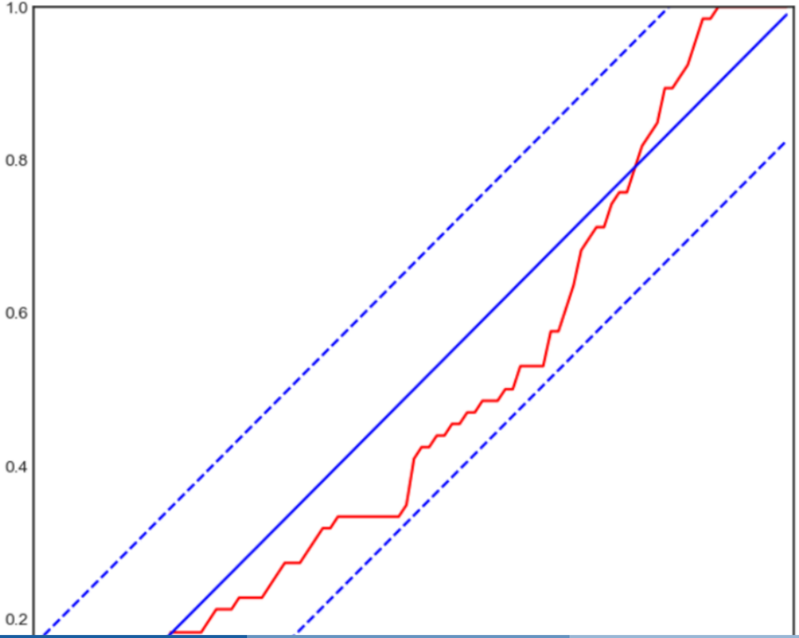
Intuition

Intuitively, the forecasted quantiles from a correctly specified model should appear as frequently as their realizations. For instance, the true values should occur less than 10% of the 10th quantile

- **Pessimistic model**: if the true values below the forecasted 10th percentile appear significantly more than 10% of the time
- **Optimistic model**: if the true values below the forecasted 10th percentile appear significantly less than 10% of the time

- To quantify this approach, the PIT Test uses the concept of the probability integral transform
- A probability integral transform is simply the evaluation of the cdf of a random variable (F_x) on its own realizations (X_t)
- Mathematically, the random variable $Y = F_X(X)$ should be uniformly distributed and it is possible to test the deviation from the uniform distribution

Probability Integral Transform



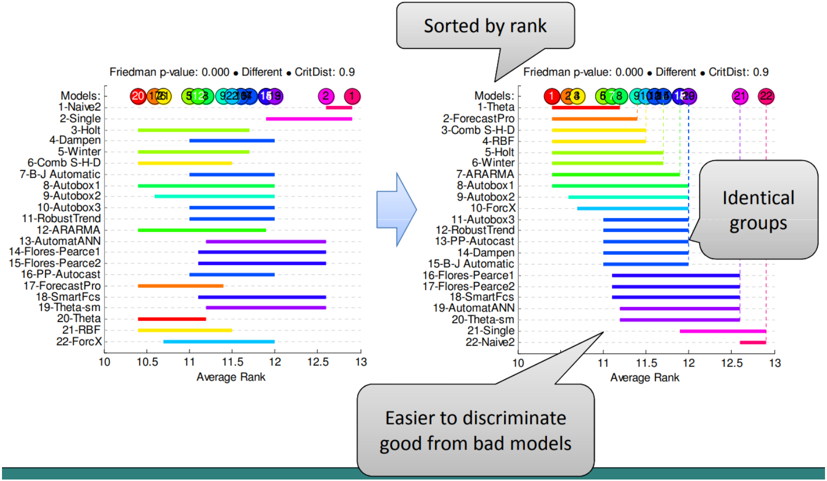
Scoring Tests

- PIT test answers the question: "is my model well specified"?
- But it doesn't inform about the performance. If two models are well specified, how can we distinguish between them?
- Idea: score them based on their *ex-post* performance of their *ex-ante* forecasts

Intuition

- Idea: what was the *ex-ante* probability of the *ex-post* realization?
- Scores are usually taken in log-form: $S\left[\hat{f}_t(y_{t+h})\right] = \log\left(\hat{f}_t(y_{t+h})\right)$

Ex-Ante Probability and Ex-Post Realizations



Source: *Lafarguette (2019)*

Tests for Equal Predictive Ability using Logscores

- A logscore is a relative metric, for a single model, it doesn't inform (at the difference of PIT tests)
- However, the difference of logscores between models informs whether a model performs better than another one and should be preferred
- Need to assess whether the difference is significant if we want to test a model $\hat{f}$ against another one $\hat{g}$
- Use the test of equal predictive ability via a Diebold-Mariano metric (1995)

Asymmetric Logscores

- The simple difference provides information about how models performs "on average"

- However, density forecasts are especially useful to inform about risks

- Hence, it makes sense to use **asymmetric logscores** to **test the performance in certain parts of the forecasted distribution**, especially the tails