Rillor

METHODS NOTE

How to tell whether a forecast beats a naive baseline

How to compare forecast accuracy against simple baselines on data the model never saw, using scaled errors, skill scores and a test that separates skill from luck.



14 November 2026 Forecasting and market intelligence rillor.com/insights/forecast-accuracy-against-naive-baselines

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

A forecast beats a naive baseline when it scores better than simple forecasts named in advance, on data none of them saw, by a margin that a test can separate from luck.

- Simple forecasts set a high bar. The naive forecast repeats the last observed value and is optimal when data follow a random walk [1]. In the M4 forecasting competition, none of the six pure machine learning (ML) methods was more accurate than the combination benchmark [14].
- The baseline has to suit the series. On quarterly beer production, the seasonal naive forecast, which repeats the value from the same quarter a year earlier, had a root mean squared error (RMSE) of 14.31, against 62.69 for the naive forecast [1].
- Percentage errors fail when an actual value is zero [1]. The mean absolute scaled error (MASE) divides each error by the in-sample error of the one-step naive forecast, so a value below 1 beats that yardstick, though multi-step forecasts typically score above 1 [2].
- A skill score reads a forecast's score against a reference forecast, with 1 for a perfect forecast and 0 for no improvement over the reference [26].
- A smaller error in one sample can be luck. The Diebold-Mariano (DM) test asks whether the mean loss difference between two forecasts is zero [8, 9]. In one small-sample case the original test rejected a true null 20.3% of the time at a nominal 10% level, against 14.2% for the Harvey-Leybourne-Newbold modification [7].
- Diebold wrote that the Diebold-Mariano test "was intended for comparing forecasts" and "was not intended for comparing models" [9]. A larger model that contains a baseline as a special case (a nested model), and a search over many models, each have tests built for them [10, 12, 13].
- In the Philadelphia Fed's Survey of Professional Forecasters (SPF) statistics, dated August 2025, current-quarter unemployment forecasts had an RMSE 0.335 times the no-change forecast's over 153 forecasts, with a Diebold-Mariano p-value of 0.165 [21].
- One forecasting benchmark calculates analytically that detecting an edge of 0.02 on its own score at 80% power takes about 350 resolved binary predictions, and an edge of 0.01 four times as many [33].

Background

A naive forecast needs no model. It repeats the last value observed. Hyndman and Athanasopoulos use it as one of four "extremely simple and surprisingly effective" benchmarks, and compare every new method with them "to ensure that the new method is better than these simple alternatives. If not, the new method is not worth considering" [1]. The naive method "works remarkably well for many economic and financial time series" and is optimal when the data follow a random walk [1]. For a random walk, Hewamalage, Ackermann and Bergmeir write, "any superiority, e.g., in error evaluations is by pure chance", and for series close to random walks, such as wind power and wind speed, the naive forecast is "a trivial yet competitive benchmark" [28].

Competitions show how often simple methods hold up. In M4, "Out Of the 17 most accurate methods, 12 were 'combinations' of mostly statistical approaches", and "The six pure ML methods performed poorly, with none of them being more accurate than the combination benchmark and only one being more accurate than Naïve2" [14]. On 1045 monthly series from the earlier M3 competition, popular machine learning methods were "dominated across both accuracy measures used and for all forecasting horizons examined" by eight statistical ones [15]. M5, on 42,840 series of Walmart unit sales [16], shows the other side. Of 5,507 teams, 48.4% beat the naive benchmark, 35.8% the seasonal naive benchmark and 7.5% the "top-performing benchmark (ES bu)", and the winning team improved on that benchmark by 22.4% [17]. Strong entries beat simple benchmarks, and most teams did not.

Macroeconomic forecasting has the same record. Atkeson and Ohanian compared three sets of inflation forecasts from NAIRU models, including the Federal Reserve Board's, with a forecast that inflation over the next year equals inflation over the last year, and found that "none of the NAIRU forecasts is more accurate than the naive forecast" [18]. As Dotsey, Fujita and Stark summarize, that study covered January 1984 to November 1999, while Stock and Watson found Phillips curve forecasts superior over 1970 to 1983 and traced the difference to "the different sample period" [19].

This note is for analysts, data scientists and planning teams who need to show that a forecast beats the obvious alternative, and for reviewers of such claims.

Method

What was examined. Each step of a comparison against simple baselines, from choosing the baselines to counting comparisons and forecasts, read from the source that defines or measures it. The Federal Reserve Bank of Philadelphia's error statistics for the SPF serve as the worked example.

Sources. Thirty-three sources, each opened on 8 November 2026 and read for the passages cited: a forecasting textbook chapter, 12 journal articles, 8 working papers and preprints, 4 software documentation pages and source files, 7 Federal Reserve publications and data files, and a forecast verification guide.

How. Figures are taken as each source publishes them. Figures marked "derived" are arithmetic on a source's own numbers or formula. SPF figures use the files' initial-release realizations as outcomes [21, 22, 23, 24].

Where sources disagree. The textbook says the mean absolute percentage error (MAPE) puts "a heavier penalty on negative errors than on positive errors" [1], and Hyndman and Koehler say "a heavier penalty on positive errors than on negative errors" [2]. This note says only that it treats over- and under-forecasts unequally. scikit-learn's guide rules out shuffled K-fold splits on time series [31], while Bergmeir, Hyndman and Koo show that standard K-fold cross-validation is possible for purely autoregressive models with uncorrelated errors [29]. This note treats time-ordered splits as the general case.

Findings

1. Fix the baselines first: naive, seasonal naive, drift and historical mean

Four benchmarks. The textbook's mean method sets every forecast to "the average (or 'mean') of the historical data", the naive method sets them "to be the value of the last observation", the seasonal naive method uses "the last observed value from the same season (e.g., the same month of the previous year)", and the drift method is "equivalent to drawing a line between the first and last observations, and extrapolating it into the future" [1].

Persistence and climatology. In weather verification, skill is "the relative accuracy of the forecast over some reference forecast", where the reference is generally "an unskilled forecast such as random chance, persistence ... or climatology" [26]. The WWRP/WGNE Joint Working Group adds that "random chance is usually not the best forecast to compare to", and that "it may be better to use climatology (long-term average value) or persistence" [26]. Persistence is the no-change forecast, and climatology the historical rate.

The right baseline depends on the series. The textbook scores all four on quarterly Australian beer production over a 2008 to 2010 test period [1].

Method	RMSE	MASE
Seasonal naive	14.31	0.94
Mean	38.45	2.44
Naive	62.69	4.01
Drift	64.90	4.12

Source: [1].

"All of the results point to the seasonal naïve method as the best of these four methods for this data set" [1]. Seasonal naive cuts the naive forecast's RMSE by 77% (derived) [1], so a method that beats only the naive forecast on seasonal data has beaten a weak opponent.

A weak baseline flatters. Stark found that "The survey's projections easily outperform no-change forecasts for all variables except long-term interest rates", and for real output traced the ease to "low persistence in the quarterly growth rates of real variables in the U.S. data" [25]. Against autoregressive benchmarks, beyond the shortest horizon, "it is hard to distinguish the SPF forecasts for real output from the benchmark projections" [25].

Same information, same moment. SPF panelists "have access to some monthly indicators for the first month of each quarter", which puts them at an advantage over the benchmarks for some variables [21]. A fourth benchmark adds monthly information, and for interest rates and the unemployment rate "the accuracy of the benchmark DARM projections rivals that of the SPF projections" [21]. Each benchmark is estimated as if in real time, on "a rolling, fixed window of 60 real-time quarterly observations" with "the vintage of data that the survey panelists would have had", except for CPI inflation, whose benchmarks use the mid-quarter vintage, one month later [21].

A fill-in baseline in an LLM benchmark. ForecastBench imputes a forecast when a model skips a question: "the value of the crowd forecast on the forecast due date" on market questions, and a value that "represents an uninformed 50% forecast" on dataset questions [32]. This "Imputed Forecaster" ranked 17th of the 50 entries in the paper's table for the LLM question set, with a Brier score of 0.149 against 0.123 for the top entry, so 33 entries ranked below it (derived from its Table 19) [32]. It copies the crowd on market questions, where it scored 0.048, so it is not a pure naive forecast [32].

2. Hold out data and use a rolling origin

New data only. "The accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when fitting the model" [1]. The test set is "typically about 20% of the total sample" and "should ideally be at least as large as the maximum forecast horizon required" [1]. "A perfect fit can always be obtained by using a model with enough parameters" [1].

Rolling origin. Each test set is a single observation, and "The corresponding training set consists only of observations that occurred prior to the observation that forms the test set", so the origin of the forecast "rolls forward in time" [1]. In the textbook's example of the drift method on a daily stock price series from 2015, the cross-validated RMSE was 11.27 against 11.15 from the training residuals, because those residual "forecasts" come from a model fitted to the entire data set, "rather than being true forecasts" [1].

Time-ordered splits. scikit-learn's TimeSeriesSplit exists because other cross-validation methods "would lead to training on future data and evaluating on past data", and its `gap` parameter excludes samples "from the end of each train set before the test set" [30]. The user guide notes that K-fold and shuffle splits "assume the samples are independent and identically distributed" [31]. The exception is narrow: for "a purely autoregressive model, the use of standard K-fold CV is possible as long as the models considered have uncorrelated errors" [29].

Leakage and vintages. Leakage is "the inadvertent use of data from the test set, or more generally data not available during inference, while training a model" [28]. Diebold calls for "period-by-period 'vintage' data" in real-time comparisons, which is "tremendously relevant for variables subject to revision, such as most macroeconomic series" [9]. Stark found that revisions "often affect absolute measures of accuracy" but "tend not to change the survey's accuracy relative to the benchmarks" [25].

Splits can be mined. Out-of-sample methods "actually expand the scope for data mining in finite samples ... because one can also mine over the sample split point" [9]. The split is fixed before any results are seen.

3. Prefer scaled errors to percentage errors

Scale and target. Mean absolute error (MAE) and RMSE "cannot be used to make comparisons between series that involve different units", and minimizing MAE leads to forecasts of the median while minimizing RMSE leads to forecasts of the mean [1].

Percentage errors. MAPE is "infinite or undefined if y_t = 0 for any t in the period of interest", needs a meaningful zero, which rules out temperatures in Fahrenheit or Celsius, and treats over- and under-forecasts unequally [1, 2]. "Hyndman & Koehler (2006) recommend that the sMAPE not be used" [1].

MASE. Hyndman and Koehler found the measures used in the M and M3 competitions "degenerate in commonly occurring situations" and proposed "that the mean absolute scaled error become the standard measure for comparing forecast accuracy across multiple time series" [2]. Each error is divided by the in-sample mean absolute error of the one-step naive forecast, and "A scaled error is less than one if it arises from a better forecast than the average one-step naïve forecast computed in-sample" [2]. It is infinite or undefined only "when all historical observations are equal" [2]. The textbook gives a seasonal version, scaled by seasonal naive errors, and the root mean squared scaled error (RMSSE) [1].

The multi-step trap. "Because the scaling is based on one-step forecasts, the scaled errors for multi-step forecasts are typically larger than one" [2]. A MASE above 1 at a long horizon does not show a loss to the naive forecast [2]. The test of that is the naive method's own out-of-sample errors on the same origins and horizons.

Scaled errors fit the test. Franses writes that "the MASE fits nicely within the standard statistical procedures initiated by Diebold and Mariano (1995)" [3]. M5 used a weighted RMSSE, since "measures based on scaled errors are probably the ones that display the most preferable statistical properties", though "no measure is perfect" [17].

Match the score to the forecast. Gneiting warns that averaging absolute or squared errors "can lead to grossly misguided inferences, unless the scoring function and the forecasting task are carefully matched". The scoring function should "be specified ex ante", or the forecaster told which statistical functional to report, such as the mean or a quantile [4].

No universal level. None of the sources reviewed gives a general good value for MASE, MAPE or percent accuracy. Some forecasts "may be more accurate simply because the weather is easier to forecast", and skill takes this into account [26].

4. Report skill against the baseline

The general form. A skill score is (score of the forecast - score of the reference) / (score of a perfect forecast - score of the reference). Its upper bound is 1, and 0 means no improvement over the reference [26]. It "can be unstable for small sample sizes", and with mean squared error as the score it is called the reduction of variance [26].

A worked example. On the continuous ranked probability score (CRPS), the textbook's skill of drift relative to naive is (CRPS of naive - CRPS of drift) / CRPS of naive. For forecasts of January 2016 from the 2015 Google price series it was -0.266: "the drift method is 26.6% worse than the naïve method" [1]. The test set must be "large enough to allow reliable calculation of the error measure, especially in the denominator", which is why "MASE or RMSSE are often preferable scale-free measures for point forecast accuracy" [1].

Correlation is not skill. Murphy splits mean squared error skill scores into terms for correlation, conditional bias and unconditional bias, and the decompositions "document fundamental deficiencies in the correlation coefficient as a measure of performance" [27].

RMSE ratios. The Philadelphia Fed reports the ratio of the SPF's RMSE to each benchmark's, and "A RMSE ratio below unity indicates that the SPF consensus (median) forecast has a root-mean-square error lower than that of the benchmark" [21]. One minus the squared ratio is the skill on mean squared error (derived) [21, 26].

Probability forecasts. The Brier skill score and its small-sample limits are covered in Brier score and log loss for probability forecasts (https://rillor.com/insights/brier-score-and-log-loss).

5. Test the gap: the Diebold-Mariano test and its small-sample correction

Why a test. In Diebold's example, two forecasts have sample mean squared errors of 1.80 and 1.92, and "the question arises in any particular sample as to whether S is truly superior or merely lucky" [9].

The test. Diebold and Mariano test "the null hypothesis of no difference in the accuracy of two competing forecasts", where "the loss function need not be quadratic and need not even be symmetric" and errors "can be non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated" [8]. The loss differential is one forecast's loss minus the other's at each date. The test "is simply an asymptotic z-test of the hypothesis that the mean of an observed series (the loss differential) is zero", with a standard error calculated robustly, for example by regressing the loss differential on an intercept with heteroskedasticity and autocorrelation robust (HAC) standard errors [9]. "DM requires only that the loss differential be covariance stationary", which Diebold suggests checking by plotting the series and its sample autocorrelations [9]. For optimal forecasts h steps ahead, errors follow "a moving average process of order (h - 1)" [7].

The small-sample correction. Harvey, Leybourne and Newbold found that the original test "can be quite seriously over-sized in small and moderate samples" [7]. Their modification adjusts the statistic and uses "critical values of the Student's t rather than the standard normal distribution" [7]. With two-step forecasts, 16 observations and a nominal 10% level, "The original test rejects the null 20.3% of the time, while the comparable figure for the modified test is 14.2%" [7]. The authors say it "goes a long way towards rectifying this problem, but does not cure it entirely", and that it applies to "cost-of-error functions other than quadratic loss" [7].

As implemented. The forecast package's dm.test "implements the modified test proposed by Harvey, Leybourne and Newbold (1997)", with arguments for the alternative hypothesis, the horizon h and the power of the loss, "Usually 1 or 2" [5]. The code multiplies the statistic by k = ((n + 1 - 2h + (h/n)(h - 1))/n)^(1/2) and uses the t distribution with n - 1 degrees of freedom [6].

Forecasts (n)	Horizon (h)	Scaling factor k
20	1	0.975
20	4	0.825
50	4	0.930
100	4	0.965

Source: derived from the dm.test code [6].

In public use. The Philadelphia Fed reports three two-sided p-values for each SPF and benchmark pair: the DM statistic, "The Harvey-Leybourne-Newbold correction (1997, International Journal of Forecasting) to the Diebold-Mariano statistic" and DM with a longer Bartlett window [21].

6. What the test was not built for

Forecasts, not models. In his retrospective (NBER working paper, September 2012, published in 2015), Diebold wrote that the DM test "was intended for comparing forecasts" and "was not intended for comparing models" [9]. When models are compared out of sample, "much simpler yet more compelling full-sample model comparison procedures exist" [9]. He finds the hunch that out-of-sample analysis is even a good insurance against in-sample over-fitting in model comparisons "largely false", while it "may be useful for learning about comparative historical predictive performance" [9].

Forecast errors as primitives. The test takes "forecast errors as primitives, intentionally. Those forecast errors need not come from models", and Diebold lists forecasts from surveys, financial markets, prediction markets and expert judgment [9]. Forecasts logged by LLM agents and other superintelligence (SI) systems enter the test the same way, as errors against outcomes.

Nested models. When a larger model nests a parsimonious one, "the larger model introduces noise into its forecasts by estimating parameters whose population values are zero", so the parsimonious model's mean squared prediction error "is therefore expected to be smaller", and Clark and West "describe how to adjust MSPEs to account for this noise" [10]. A model that adds predictors to a naive or autoregressive baseline nests it.

Conditional tests. Giacomini and White's tests allow for a "possibly misspecified" forecasting model, "capture the effect of estimation uncertainty on relative forecast performance" and "can handle forecasts based on both nested and nonnested models" [11].

A minefield. For tests that compare models out of sample, Diebold warns that "one must tiptoe carefully across a minefield of assumptions", including whether the estimation sample is "fixed? Recursively expanding? Rolling?" [9].

7. A public example: professional forecasts against no-change benchmarks

The record. The Philadelphia Fed compares the SPF median forecast with four benchmarks: indirect autoregressive (IAR), no-change, direct autoregressive (DAR) and direct autoregressive with monthly information (DARM) [20]. The no-change forecast "is simply the historical value for the prior quarter", "perhaps the simplest of all possible benchmarks" [21]. Horizon 1 is the current quarter and horizon 5 is four quarters ahead, and sample periods "refer to the dates forecast, not the dates when the forecasts were made" [21]. The text files were released on 22 August 2025 [21, 22, 23, 24] and the page is dated 25 August 2025 [20]. It says that "Due to complexities introduced by the 2025 federal government shutdown, we have paused updates of the SPF Error Statistics" [20].

Variable, target quarters	Forecasts	Horizon	Benchmark	RMSE ratio	DM p-value
Unemployment, 1985 Q1 to 2023 Q1	153	1	No change	0.335	0.165
Unemployment, 1985 Q1 to 2023 Q1	153	1	IAR	0.168	0.058
Unemployment, 1985 Q1 to 2023 Q1	153	5	No change	0.765	0.065
Unemployment, 1985 Q1 to 1996 Q4	48	1	No change	0.635	0.001
Unemployment, 1985 Q1 to 1996 Q4	48	1	DARM	0.965	0.760
Real GDP growth, 1985 Q1 to 2023 Q1	151	4	IAR	0.988	0.316
CPI inflation, 1997 Q1 to 2023 Q1	105	1	No change	0.533	0.011
CPI inflation, 1997 Q1 to 2023 Q1	105	5	DAR	0.957	0.023
10-year Treasury rate, 1993 Q1 to 2023 Q1	121	1	No change	0.450	0.000

Variable, target quarters	Forecasts	Horizon	Benchmark	RMSE ratio	DM p-value
10-year Treasury rate, 1993 Q1 to 2023 Q1	121	4	No change	1.025	0.695
10-year Treasury rate, 1993 Q1 to 2023 Q1	121	5	No change	1.075	0.332

Sources: unemployment [21], real GDP [23], CPI [22], 10-year Treasury rate [24]. Outcomes are first-release values (for unemployment and CPI, the mid-quarter vintage, which reflects some revision) [21].

A large ratio is not a significant one. Over the long sample, current-quarter unemployment forecasts had about a third of the no-change forecast's RMSE, with p-values of 0.165 (DM) and 0.168 (Harvey-Leybourne-Newbold) [21]. Over 1985 to 1996, a ratio of 0.635 gave 0.001 and 0.002 [21]. The survey's own RMSE at horizon 1 was 0.14 percentage points over 1985 to 1996 and 0.35 over 1997 to 2023 [21].

Small gaps against autoregressive benchmarks. For real GDP growth at horizon 4, a ratio of 0.988 against the indirect autoregressive benchmark is an MSE skill of about 0.024 (derived), with a DM p-value of 0.316 [23]. That fits Stark's finding for real output beyond the shortest horizon [25]. For CPI inflation four quarters ahead, a ratio of 0.957 against the direct autoregressive benchmark gave p-values of 0.023 (DM) and 0.032 (Harvey-Leybourne-Newbold) [22].

The baseline can win. For the 10-year Treasury rate at horizons 4 and 5, the no-change forecast had the lower RMSE, though neither difference was significant [24]. Stark found that "For predicting long-term interest rates at the longest horizons, the no-change benchmark performs as well as the SPF and better than the remaining benchmarks" [25]. These are forecasts of an interest rate level, read here for accuracy only.

Horizon by horizon. "The survey's accuracy falls sharply at quarterly horizons beyond the first", and "At long horizons, relative accuracy depends on the variable forecast and the benchmark model considered" [25].

8. Count every comparison, and collect enough forecasts

Data snooping. "Data snooping occurs when a given set of data is used more than once for purposes of inference or model selection", and with time series it "is practically unavoidable", since "typically only a single history measuring a given phenomenon of interest is available" [12]. White's reality check tests "the null hypothesis that the best model encountered in a specification search has no predictive superiority over a given benchmark model" [12].

Model confidence sets. Hansen, Lunde and Nason define the model confidence set as "a set of models that is constructed such that it will contain the best model with a given level of confidence", "analogous to a confidence interval for a parameter", and uninformative data yield a set with many models [13]. Diebold notes that robustness to data mining in finite samples "requires simulation methods", such as White's reality check or the bootstrap model confidence set [9].

Many methods at once. Hewamalage and colleagues describe the Friedman test on ranks, usually followed by a post-hoc procedure such as Holm, Nemenyi or Multiple Comparisons with the Best, and rate a comparison with the mean lower, "since it is more useful to find which methods are not significantly different from the best, than from some averagely performing method overall" [28].

Selection on validation scores. In M5, 48.4% of teams beat the naive benchmark with the submissions they selected, while 63.7% had a submission that would have, and the share with one that beat ES bu was 12.2% against 7.5% [17]. The authors conclude that "many teams failed to choose the best method developed, probably due to misleading validation scores" [17].

How many forecasts. Foresight Arena, a benchmark for forecasting agents, calculates that "detecting a true edge of alpha* = 0.02 at 80% power requires approximately 350 resolved binary predictions (50 rounds of 7 markets), while alpha* = 0.01 requires four times more" [33]. The figure is analytical, on its own Alpha score, an edge over market consensus, from "standard normal theory" under stated assumptions [33]. The verification guide calls error bounds "especially important" for rare events, variable data and claims that one forecast is significantly better than another, and warns that pooling "can mask variations in forecast performance when the data are not homogeneous" [26].

Implications

For teams building a forecast. A comparison that holds up:

1. Names the baselines and the error measure before the first forecast is scored, choosing among naive, seasonal naive, drift and the historical mean to suit the series [1, 4, 9].
2. Gives the baselines the same data, vintages and origins as the forecast [9, 21].
3. Holds out a test set at least as long as the longest horizon, uses a rolling origin and fixes the split in advance [1, 9].
4. Reports scaled errors or a skill score against the named baseline, horizon by horizon [1, 2, 26].
5. Tests the gap with the modified DM test after checking the loss differential's autocorrelation, and uses the tests built for nested models or model searches where they apply [7, 9, 10, 12].
6. Reports how many models, horizons and splits were tried, and corrects for them [12, 13].
7. Collects enough resolved forecasts to separate the expected edge from luck [7, 33].

For reviewers of an accuracy claim. Ask which baselines were named and when, whether they had the same information, how the test period was chosen, which error measure was used, whether the gap was tested, and how many comparisons sit behind the headline figure.

At Rillor. Rillor's forecasting and market intelligence work compares each forecast with simple baselines, such as no change or the historical rate. The same pipeline also runs on randomly chosen dates, and a result counts only when it separates from those controls.

Limits of this analysis

- The SPF figures end with target quarter 2023 Q1 [21, 22, 23, 24]. The files are dated 22 to 25 August 2025, and updates are paused [20, 21].
- M4 is reported only as its abstract words it [14]. The M5 figures are from the October 2020 preprint [17].
- Foresight Arena's sample-size figure is analytical, on its own Alpha score and under its stated assumptions, and its version 2 note calls its section 6 "an illustrative simulation study" [33].
- ForecastBench's imputed forecaster copies the crowd on market questions, and its figures are from the paper's version of 28 February 2025 [32].
- A verdict against a baseline can change with the sample period [19, 21].
- The DM test is asymptotic, and the Harvey-Leybourne-Newbold modification does not cure its excess size entirely [7, 9].
- This note describes evaluation methods from published sources. It is research, never investment advice, and it ranks no forecaster, model family or product.

Sources

1. Forecasting: Principles and Practice, third edition, chapter 5, The forecaster's toolbox. Rob J Hyndman and George Athanasopoulos, OTexts, site modified 4 November 2026, viewed 8 November 2026. https://otexts.com/fpp3/toolbox.html

2. Another look at measures of forecast accuracy. Rob J Hyndman and Anne B Koehler, author copy dated 2 November 2005 (International Journal of Forecasting, 2006), viewed 8 November 2026. https://robjhyndman.com/papers/mase.pdf

3. A note on the Mean Absolute Scaled Error. Philip Hans Franses, International Journal of Forecasting 32(1), 20-22, 2016, IDEAS/RePEc record, viewed 8 November 2026. https://ideas.repec.org/a/eee/intfor/v32y2016i1p20-22.html

4. Making and Evaluating Point Forecasts. Tilmann Gneiting, arXiv 0912.0902, version 2, 7 March 2010, viewed 8 November 2026. https://arxiv.org/abs/0912.0902

5. dm.test: Diebold-Mariano test for predictive accuracy. George Athanasopoulos and Kirill Kuroptev, forecast R package reference, version 9.0.2.9000, viewed 8 November 2026. https://pkg.robjhyndman.com/forecast/reference/dm.test.html

6. R/DM2.R, source of dm.test in the forecast package. Rob J Hyndman and contributors, GitHub, master branch, viewed 8 November 2026. https://raw.githubusercontent.com/robjhyndman/forecast/master/R/DM2.R

7. Testing the equality of prediction mean squared errors. David Harvey, Stephen Leybourne and Paul Newbold, International Journal of Forecasting 13(2), 281-291, June 1997, university-hosted copy, viewed 8 November 2026. http://eurequa.univ-paris1.fr/membres/Ahamada/cours/article3.pdf

8. Comparing Predictive Accuracy. Francis X Diebold and Roberto S Mariano, Journal of Business & Economic Statistics 13(3), 253-263, July 1995, IDEAS/RePEc record, viewed 8 November 2026. https://ideas.repec.org/a/bes/jnlbes/v13y1995i3p253-63.html

9. Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests. Francis X Diebold, NBER Working Paper 18391, September 2012 (Journal of Business & Economic Statistics, 2015), National Bureau of Economic Research, viewed 8 November 2026. https://www.nber.org/papers/w18391

10. Approximately normal tests for equal predictive accuracy in nested models. Todd E Clark and Kenneth D West, Journal of Econometrics 138(1), 291-311, May 2007, IDEAS/RePEc record, viewed 8 November 2026. https://ideas.repec.org/a/eee/econom/v138y2007i1p291-311.html

11. Tests of Conditional Predictive Ability. Raffaella Giacomini and Halbert White, Econometrica 74(6), 1545-1578, November 2006, IDEAS/RePEc record, viewed 8 November 2026. https://ideas.repec.org/a/ecm/emetrp/v74y2006i6p1545-1578.html

12. A Reality Check for Data Snooping. Halbert White, Econometrica 68(5), 1097-1126, September 2000, The Econometric Society, viewed 8 November 2026. https://www.econometricsociety.org/publications/econometrica/2000/09/01/reality-check-data-snooping

13. The Model Confidence Set. Peter R Hansen, Asger Lunde and James M Nason, Econometrica 79(2), 453-497, March 2011, IDEAS/RePEc record, viewed 8 November 2026. https://ideas.repec.org/a/ecm/emetrp/v79y2011i2p453-497.html

14. The M4 Competition: Results, findings, conclusion and way forward. Spyros Makridakis, Evangelos Spiliotis and Vassilios Assimakopoulos, International Journal of Forecasting 34(4), 802-808, 2018, IDEAS/RePEc record, viewed 8 November 2026. https://ideas.repec.org/a/eee/intfor/v34y2018i4p802-808.html

15. Statistical and Machine Learning forecasting methods: Concerns and ways forward. Spyros Makridakis, Evangelos Spiliotis and Vassilios Assimakopoulos, PLOS ONE 13(3): e0194889, 27 March 2018, viewed 8 November 2026. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0194889

16. M5 accuracy competition: Results, findings, and conclusions. Spyros Makridakis, Evangelos Spiliotis and Vassilios Assimakopoulos, International Journal of Forecasting 38(4), 1346-1364, 2022, IDEAS/RePEc record, viewed 8 November 2026. https://ideas.repec.org/a/eee/intfor/v38y2022i4p1346-1364.html

17. The M5 Accuracy competition: Results, findings and conclusions. Spyros Makridakis, Evangelos Spiliotis and Vassilis Assimakopoulos, preprint, October 2020, viewed 8 November 2026. https://statmodeling.stat.columbia.edu/wp-content/uploads/2021/10/M5_accuracy_competition.pdf

18. Are Phillips Curves Useful for Forecasting Inflation? Andrew Atkeson and Lee E Ohanian, Federal Reserve Bank of Minneapolis Quarterly Review 25(1), Winter 2001, Federal Reserve Bank of Minneapolis, viewed 8 November 2026. https://researchdatabase.minneapolisfed.org/concern/publications/g732d916c

19. Do Phillips Curves Conditionally Help to Forecast Inflation? Michael Dotsey, Shigeru Fujita and Tom Stark, Federal Reserve Bank of Philadelphia Working Paper 17-26, August 2017, Board of Governors of the Federal Reserve System, viewed 8 November 2026. https://www.federalreserve.gov/monetarypolicy/files/FOMC20170801memo01.pdf

20. Forecast Error Statistics for the Survey of Professional Forecasters. Federal Reserve Bank of Philadelphia, 25 August 2025, viewed 8 November 2026. https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/error-statistics

21. Error Statistics for the Survey of Professional Forecasters for Unemployment Rate. Tom Stark, Federal Reserve Bank of Philadelphia, 22 August 2025, viewed 8 November 2026. https://www.philadelphiafed.org/-/media/FRBP/Assets/Surveys-And-Data/survey-of-professional-forecasters/data-files/UNEMP/SPF_Error_Statistics_UNEMP_1_AICt.txt?sc_lang=en&hash=C815A374A57C96B9E85DF4183BDBD68A

22. Error Statistics for the SPF: CPI Inflation Rate. Tom Stark, Federal Reserve Bank of Philadelphia, 22 August 2025, viewed 8 November 2026. https://www.philadelphiafed.org/-/media/FRBP/Assets/Surveys-And-Data/survey-of-professional-forecasters/data-files/CPI/SPF_Error_Statistics_CPI_1_AICt.txt?sc_lang=en&hash=742C096200408AE41AE27D24F4EFEFD0

23. Error Statistics for the SPF: Real GNP/GDP. Tom Stark, Federal Reserve Bank of Philadelphia, 22 August 2025, viewed 8 November 2026. https://www.philadelphiafed.org/-/media/FRBP/Assets/Surveys-And-Data/survey-of-professional-forecasters/data-files/RGDP/SPF_Error_Statistics_RGDP_3_AICt.txt?sc_lang=en&hash=9B49B8FEE73D28CAD585B05436E71DDB

24. Error Statistics for the SPF: 10-Year Treasury Bond Rate. Tom Stark, Federal Reserve Bank of Philadelphia, 22 August 2025, viewed 8 November 2026. https://www.philadelphiafed.org/-/media/FRBP/Assets/Surveys-And-Data/survey-of-professional-forecasters/data-files/TBOND/SPF_Error_Statistics_TBOND_1_AICt.txt?sc_lang=en&hash=F291344796973E65B3D076CC6AA0F422

25. Realistic Evaluation of Real-Time Forecasts in the Survey of Professional Forecasters. Tom Stark, Federal Reserve Bank of Philadelphia, Research Rap special report, 28 May 2010, viewed 8 November 2026. https://www.philadelphiafed.org/-/media/FRBP/Assets/Economy/Reports/research-rap/2010/realistic-evaluation-of-real-time-forecasts.pdf?sc_lang=en&hash=A99963F337C8CE022631AEA056290D5F

26. Forecast verification - methods, issues and FAQ. WWRP/WGNE Joint Working Group on Forecast Verification Research, viewed 8 November 2026. https://jwgfvr.github.io/forecastverification/

27. Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. Allan H Murphy, Monthly Weather Review 116(12), 2417-2424, December 1988, American Meteorological Society, viewed 8 November 2026. https://journals.ametsoc.org/view/journals/mwre/116/12/1520-0493_1988_116_2417_ssbotm_2_0_co_2.xml

28. Forecast Evaluation for Data Scientists: Common Pitfalls and Best Practices. Hansika Hewamalage, Klaus Ackermann and Christoph Bergmeir, arXiv 2203.10716, version 2, 4 April 2022, viewed 8 November 2026. https://arxiv.org/abs/2203.10716

29. A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction. Christoph Bergmeir, Rob J Hyndman and Bonsoo Koo, author preprint dated 23 July 2017 (Computational Statistics and Data Analysis, 2018), viewed 8 November 2026. https://robjhyndman.com/papers/cv-wp.pdf

30. sklearn.model_selection.TimeSeriesSplit, scikit-learn 1.9.1 API reference. scikit-learn developers, viewed 8 November 2026. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html

31. Cross-validation: evaluating estimator performance, section 3.1.2.6, Cross validation of time series data, scikit-learn 1.9.1 User Guide. scikit-learn developers, viewed 8 November 2026. https://scikit-learn.org/stable/modules/cross_validation.html

32. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. Ezra Karger and others, arXiv 2409.19839, version 5, 28 February 2025, viewed 8 November 2026. https://arxiv.org/abs/2409.19839

33. Foresight Arena: An On-Chain Benchmark for Evaluating AI Forecasting Agents. Maksym Nechepurenko and Pavel Shuvalov, arXiv 2605.00420, version 2, 4 May 2026, viewed 8 November 2026. https://arxiv.org/abs/2605.00420

About this report

Published by Rillor on 14 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. How to tell whether a forecast beats a naive baseline. 14 November 2026. https://rillor.com/insights/forecast-accuracy-against-naive-baselines

Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Send the question you want forecast, its horizon and the data you already hold. Contact us at rillor.com/contact?type=forecasting or contact@rillor.com. More on forecasting and market intelligence: rillor.com/markets/forecasting