SUPPLY CHAIN

The *AI forecast accuracy* number is lying to you

When a vendor says their AI improved accuracy by 30%, three follow-up questions usually dissolve the claim: at what aggregation level, at what lag, and measured how?

MAY 22, 2026 · 4 MIN READ

When a vendor tells you their AI improved forecast accuracy by 30%, do not say "great." Ask three questions: at what aggregation level, at what forecast lag, and measured with which metric. The headline number almost always dissolves under one of them.

Start with the metric itself. Mean Absolute Percentage Error — MAPE, the industry's default — breaks at low volumes and is asymmetric. On intermittent-demand items where half the periods are zero, forecasting zero every period produces the lowest average error, which is operationally absurd. And MAPE and its weighted cousin disagree wildly on the same data: *one worked example shows MAPE at 36.7% while WAPE reads just 5.9% on the identical dataset, because a single low-volume day with 100% error inflates the unweighted average while barely moving the volume-weighted one.* A vendor quoting "accuracy" rarely tells you which one they mean — and the choice can swing the number sixfold.

Then there is aggregation, which is where most case-study numbers come from. Forecast accuracy at the SKU–DC–week level is what actually drives stockouts and excess. Accuracy at the national–monthly level is what shows up in vendor decks, because aggregation averages the errors away. A 95%-accurate national-monthly forecast routinely hides 50 to 60% error at the level where someone actually places a replenishment order.

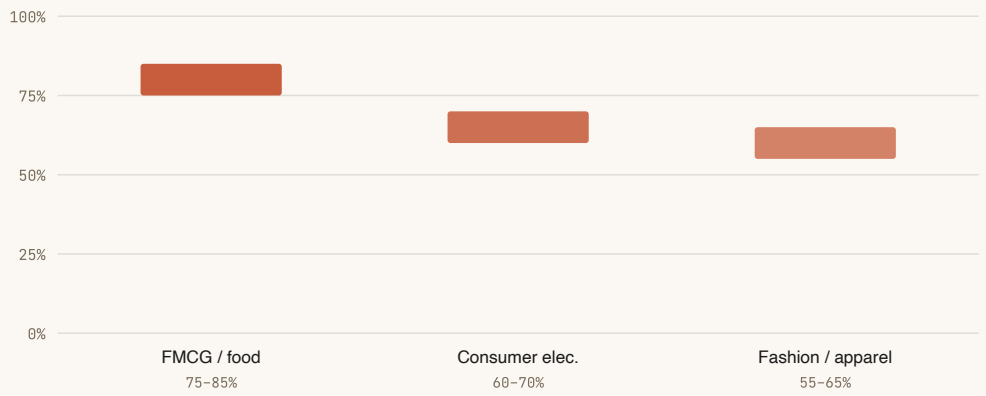
> *52%*
>
> of 300,000+ real-life forecasts were worse than a naïve random walk — the benchmark most 'accuracy' claims never disclose
>
> MORLIDGE, IBF — IN GILLILAND, TASHMAN & SGLAVO (WILEY, 2015)

And then bias, the error the headline number is structurally blind to. A forecast running +20% on half your SKUs and –20% on the other half can post a flattering aggregate MAPE while simultaneously generating stockouts on the under-forecasted items and overstock on the over-forecasted ones. Symmetric accuracy metrics cannot see this. Persistent positive bias quietly builds chronic excess; persistent negative bias builds chronic stockouts. Either way the inventory consequence is expensive and predictable — and the accuracy number tells you nothing about it.

The deepest problem is the missing benchmark. In a study of more than 300,000 real-life forecasts across eight companies, 52% were worse than a naïve random walk. *SAS's Mike Gilliland offers the practitioner's rule of thumb: if you can reduce forecast error by 10 to 20% versus a simple model, you are probably doing about the best you can expect.* An AI vendor claiming a 30% improvement should be made to show that

improvement *against a naïve benchmark* — not against your previous, possibly value-destroying, process.

TYPICAL FORECAST ACCURACY BY INDUSTRY (ILLUSTRATIVE)



COMPARING AI VENDOR CLAIMS ACROSS CATEGORIES IS APPLES-TO-ORANGES. SOURCE: DEMANDPLAN; ONEPINT SYNTHESIS.

So what should you measure instead? A defensible stack has five layers: weighted error at the actionable SKU–DC level, not aggregated; bias tracked by hierarchy, to catch over- and under-forecasting before it reaches inventory; Forecast Value Added against a naïve benchmark at *every* process step — engine, analyst, consensus, executive; and the actual outcomes, service level and inventory turns, because accuracy is a means, not an end. A probabilistic view of the forecast sits underneath all of it, because planning decisions are made under uncertainty, not against a single number.

> *Accuracy is a means, not an end. Measure the inventory and the service level — that's what the forecast was for.*
>
> ZULOMA — SUPPLY CHAIN

When the AI vendor claims 30%, the right reply is not enthusiasm. It is: show me Forecast Value Added against a random walk at SKU–DC–week, bias by hierarchy, and the inventory and service-level outcome across four full planning cycles. That is the metric set that reveals whether the AI is adding value — or just being measured kindly.

Sources

- Morlidge, S. — cited in Gilliland, Tashman & Sglavo, *Business Forecasting: Practical Problems and Solutions* (Wiley, 2015).
- Gilliland, M. (2013). *When do you stop trying to improve forecast accuracy?* SAS Blogs.
- Gilliland, M. *Forecast Value Added Analysis: Why and How.* Institute of Business Forecasting (forecasters.org).
- Prospeo. (2026). *Forecast Accuracy Metrics: Practitioner Guide.* prospeo.io
- DemandPlan; OnePint. *Forecast accuracy metrics and benchmarks.* demandplan.io; onepint.ai

---