Forecasting the budget

Every finance team forecasts the same way.
On real data, that way is wrong 10% of the time.

Kevin Larretche
Data Science × FP&A · August 2026

Project context

Rossmann, a German drugstore chain. 1,115 shops, daily sales from 1 January 2013 to 31 July 2015, 1,017,209 shop-days. A real company's real numbers, published in full.

The question. Finance has to put a sales number in the budget. Almost everyone builds it the same way: the same week last year, times a growth rate. Is that good enough, and if not, what should replace it?

The answer. That rule is wrong by 10.09% on the chain total. The method built here gets it to 3.29%. The single biggest fix costs nothing: no model, no new data, no new system.

My role

Data Science × FP&A

Backtest design: a rolling test across 49 weeks so no result depends on where you happen to stand.

Diagnosis: finding why the standard rule breaks, not just that it breaks.

Modelling: eleven methods from spreadsheet rules to daily gradient boosting, all scored the same way.

The answer: a forecast for the next two months, with a range and a named assumption.

Source: Rossmann Store Sales, public dataset, verified against the published shape before use: 1,017,209 rows, 1,115 shops, exact date range.

How every method was tested

A forecast that has only been checked once has not been checked. Every method here is run the same way, on the same weeks, against the same targets.

The method never sees the answer first

The method learns from history up to a point in time, forecasts the next six weeks, and only then are the real numbers revealed. Then everything moves forward one week and repeats.

Days are added up into weeks ending Sunday

Finance plans in weeks and months. Days are added up into weeks ending Sunday. A week counts only if all seven days are there: 133 complete weeks, 867 complete shops.

49 weeks tested, six weeks ahead each time

49 target weeks, from 24 August 2014 to 26 July 2015, six weeks ahead each time. That is 254,898 forecasts per method, and it deliberately includes four December weeks.

Scored as error, as a share of sales

Average error as a share of sales. Predict 100, sell 93, that is 7% wrong. Scored twice: per shop, and on the chain total, which is the number that goes in the board pack.

Every horizon is scored on the same target weeks, so a one-week-ahead forecast and a six-week-ahead forecast are compared on identical targets.

The eleven methods we tested

Eleven methods in four tiers, from what a finance team already does to what a competition winner would build. All on the same 49 weeks.

Tier 1 · rules you can build in Excel

Last week again. The average of the last four. The same week last year. And that times a growth rate, which is the budget rule, and what most companies run.

```
budget = sales[t − 52 weeks] × growth
```

Tier 2 · the same rule, corrected

The same idea, against a week that is genuinely comparable: matched on promotion status, then adjusted for school and public holidays and opening days.

```
rule = sales[matched week] × growth × calendar
```

Tier 3 · classical statistics, fitted per shop

Fourier regression: a smooth yearly curve plus the shop's own calendar. Theta: the M3 competition winner. Both fitted per shop, refitted every origin.

```
y = trend + Σ sin/cos + calendar
```

Tier 4 · machine learning, one model for all shops

One model over every shop at once. Two versions: one on weeks, one on days, which keeps the weekday pattern and the spike after a closed day.

```
ratio-to-last-year, 49 features, 940k rows
```

Tier 5 is not a method of its own: it is the median of four of the above, taken shop by shop and week by week. It wins.

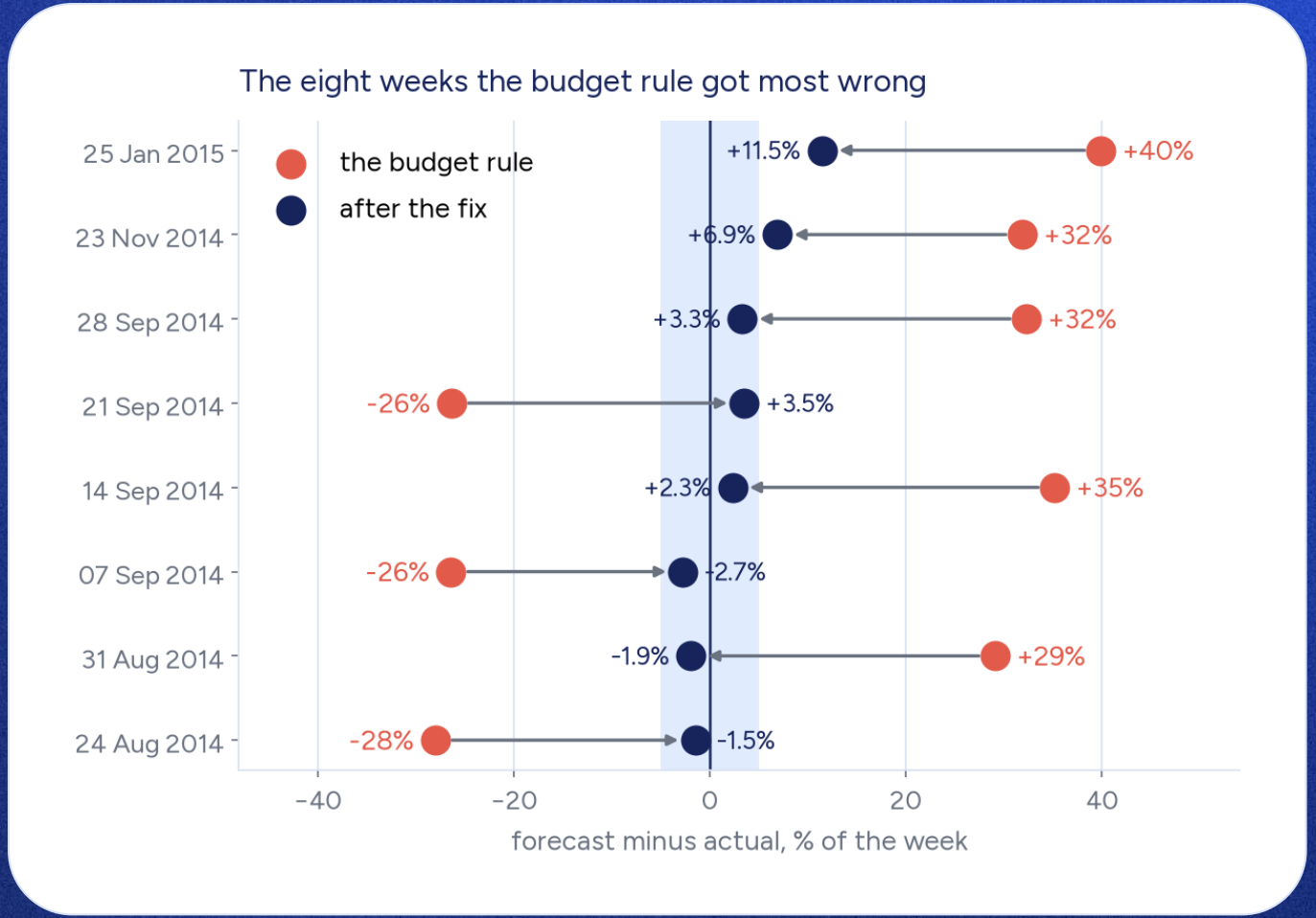
The promotion cycle breaks the budget rule

The budget rule, same week last year times growth, misses in an alternating pattern: +32%, −26%, +35%, −26%, week after week. That rhythm is a promotion cycle.

The chain promotes every second week. Mostly. The pattern breaks 13 times in two and a half years, and each break flips the rhythm permanently from then on.

So "the same week last year" lands on the same kind of week only 70% of the time. When it matches, the miss is 4.2%. When it has flipped, 29.5%.

Nobody notices, because the annual total still comes out about right.



Chain-level forecast minus actual, by week, across the 49-week test window. Labels mark the five worst weeks for the budget rule.

Matching the promotion week: 10.1% to 4.2%

Stop comparing to the calendar week. Compare to the nearest week around this time last year that had the same promotion status. A company already knows its own promotion plan, so there is nothing to buy and nothing to build.

	each shop	the chain total	what changed
the budget rule, untouched	12.62%	10.09%	the starting point
compare to the promotion-matched week	7.60%	4.19%	moves the comparison week on 16 of 49 weeks
+ growth and the calendar adjustment	7.39%	4.16%	promotion status now matches on 49 of 49

The weeks that were breaking: 25 Jan 2015 went from +39.9% to +3.8%. 14 Sep 2014 from +35.2% to 0.0%. 24 Aug 2014 from −28.0% to +2.9%. The correction leaves the other weeks alone, which is what a correction should do. School and public holidays then get the same treatment, per shop, because German states run different dates: worth a further 4.60% to 3.76% on the 23 weeks whose calendar differs from last year, and nothing on the 26 weeks where it matches.

Promotions here are chain-wide: every shop has the same promotion days in every week, always 0 or 5 days.

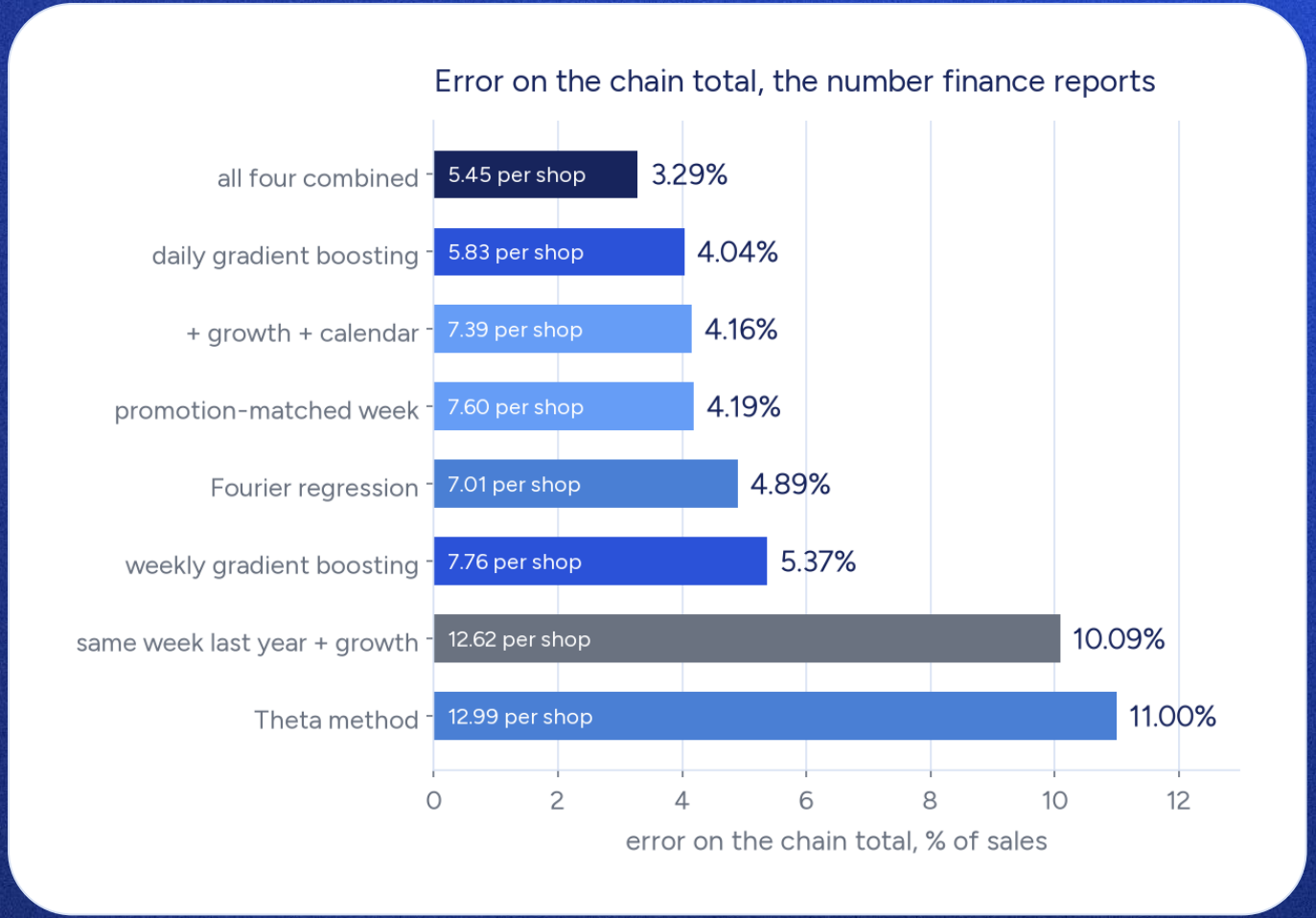
The scoreboard: every method, same 49 weeks

The best single method is daily gradient boosting, at 5.83% per shop. The best method overall is not a method at all: it is the median of four of them, taken shop by shop.

5.45% per shop, 3.29% on the chain.

Two things worth noticing. The classical statistics tier is beaten by a corrected spreadsheet rule. And combining beats every ingredient, including the ingredient that won.

Distance into the future is not the problem. The same-week-last-year rule scores 10.10% one week out and 10.10% six weeks out. Completely flat.



49 test weeks, six-week horizon, 254,898 forecasts per method. Runs made before the promotion and opening-day calendars were handled correctly are excluded.

What failed, and why

A method chapter that reports no failures is not describing real work. These were tried properly and did not earn a place.

Seasonal exponential smoothing could not be fitted at all

It needs two full yearly cycles and there are only one and a half. This is not a quirk of this dataset: it is the position every company with under two years of history is in.

Theta was the worst of the serious methods

11.00% on the chain, and 23.01% in December. It has no way to know about promotions or holidays, so it smooths straight through the weeks that matter most.

The trading-day correction over-corrects

A lost opening day costs 89% of a day, not 100%, because a public holiday moves demand rather than deleting it. Applied bluntly it repairs three weeks and breaks three others.

Reconciliation stopped earning its keep

Taking the chain total from the rule and the shop split from the model was the best trick available while the rule was broken. Once the rule was fixed, a plain median beat it: 3.29% against 4.06%.

Two mistakes of my own are recorded in the method notes: a training leak that flattered a model by 2.6 points, and a number quoted from a test window that turned out not to be representative.

Why one shop misses 5.5% but the chain 3.3%

The same forecast scores differently depending on what you add up, and the gap is not a rounding error. It decides which method you should pick.

Why the chain total is more accurate than one shop

Forecast three shops at 100. They come in at 106, 94 and 100. Per shop you were 4% wrong. On the total you were exactly right. The mistakes cancelled. With 867 shops that happens less perfectly, but it happens.

Which of the two numbers to use, and when

Putting a figure in the group budget: the chain total, 3.29%. Setting a target for one shop manager: the shop number, 5.45%. Quoting the wrong one flatters or damns a method unfairly.

Picking the wrong one picks the wrong method

The weekly boosting model is better per shop than the corrected rule, and worse on the chain. Its errors lean the same way across shops, so they add up instead of cancelling.

What to say to a shop manager on a target

A budget built shop by shop and summed is more accurate than any single shop's line in it. That is worth saying out loud to anyone being held to a shop-level target.

Error measured as absolute error over sales, the same way at both levels, so the two numbers are directly comparable.

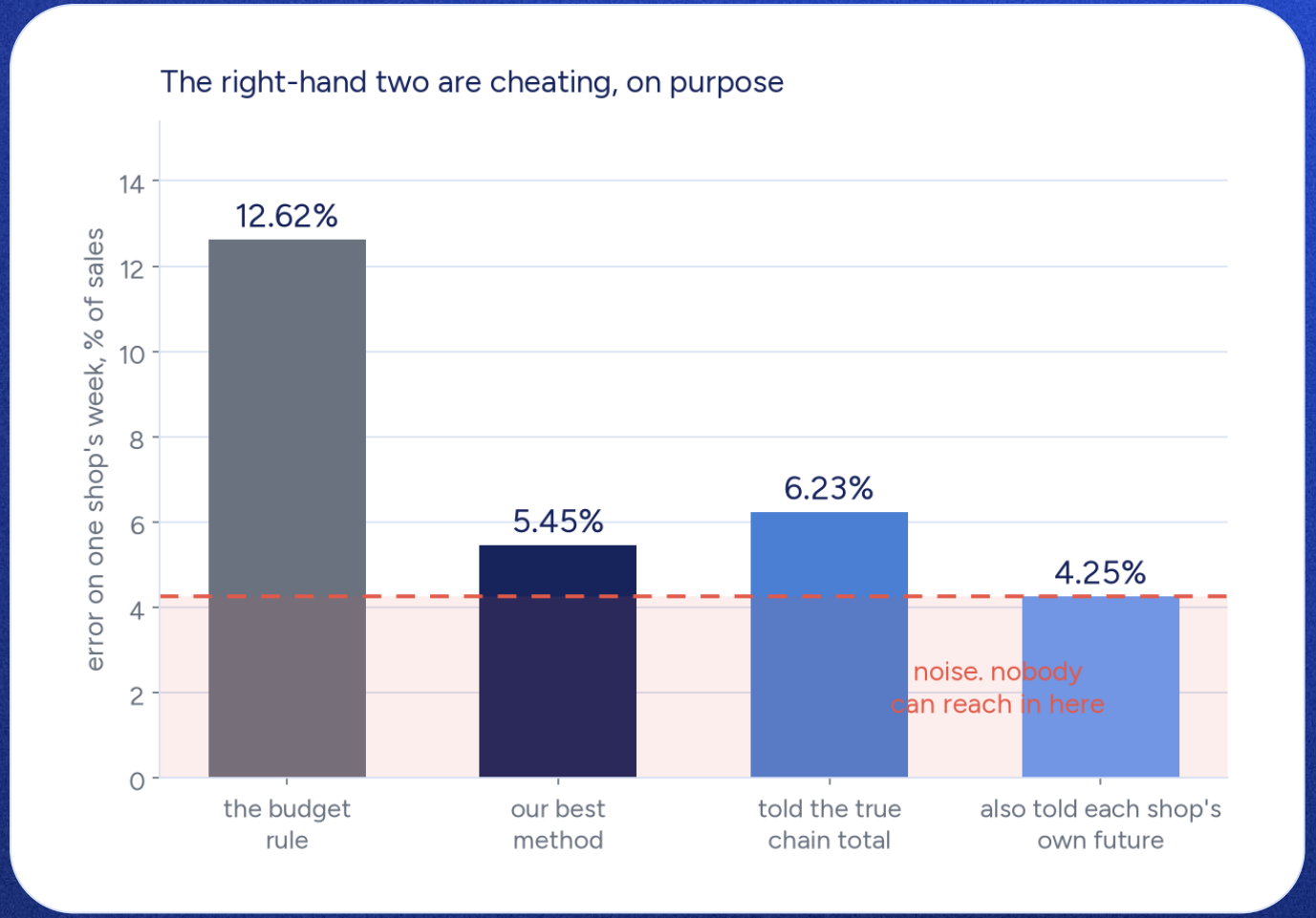
The ceiling: 4.25% per shop is unreachable

Before calling 5.45% per shop a failure, it is worth measuring the ceiling. So a forecaster was allowed to cheat.

Told the true chain total for the week in advance, each shop is still 6.23% wrong. Also told the shop's own future drift, still 4.25%.

So of our 5.45 points, 4.25 are unreachable by anyone, with any method, ever. Only about 1.2 points were ever available to win.

The leftover is local: 99% of it moves one shop on its own, not all shops together. And it is footfall, not basket. Shoppers wobble 7.6 points week to week, spend per shopper only 3.7. That is weather, roadworks, a local event.



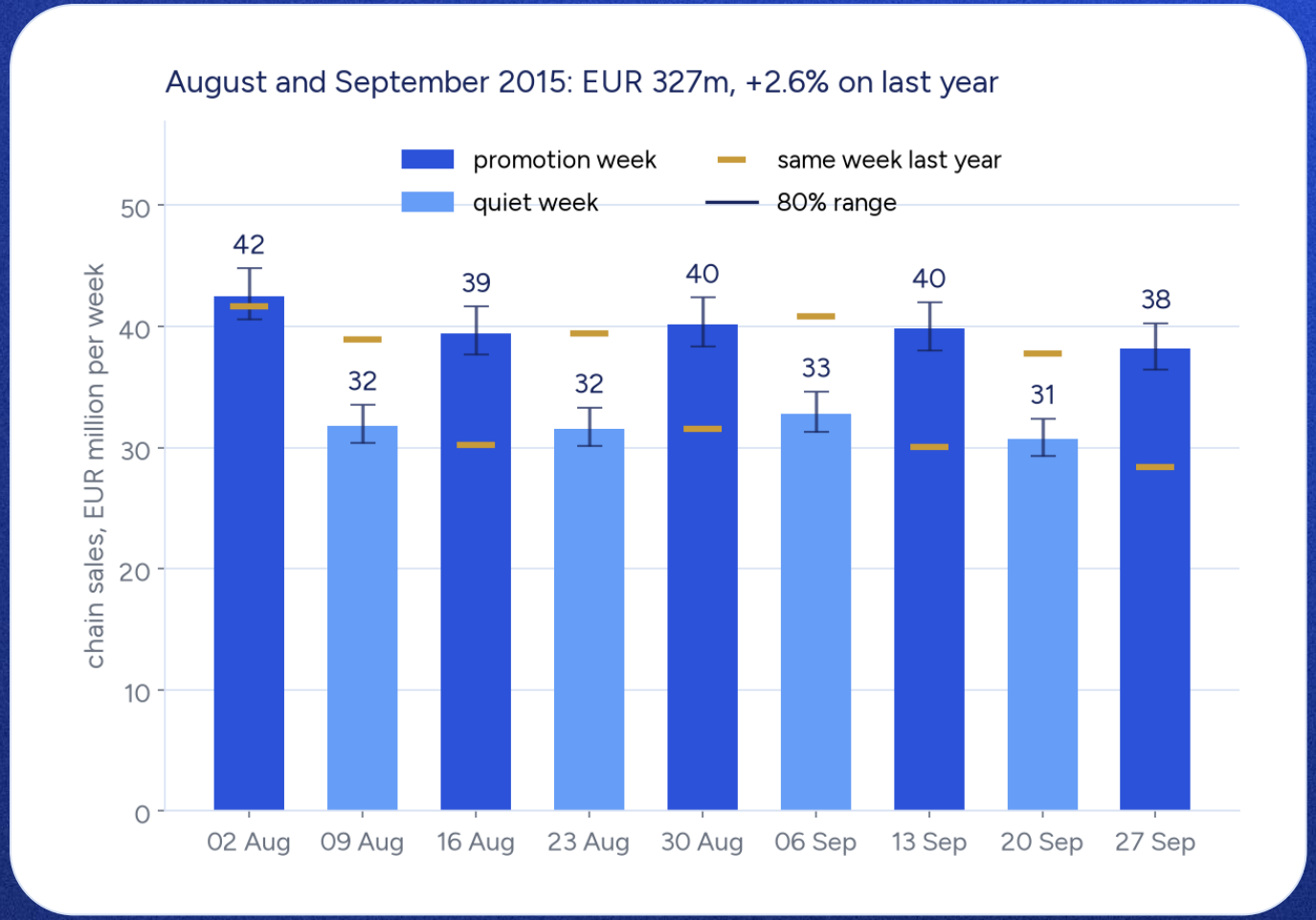
The two right-hand columns use information no forecaster could have. They exist to mark the floor, not to be beaten.

The forecast: August and September 2015

Everything so far is backtesting: scoring a method against weeks that already happened. This is the part a finance team actually wants.

The data ends Sunday 26 July 2015. Using all 133 weeks and the winning method, the nine weeks to 27 September 2015 come to EUR 327.0m, which is +2.6% on the same nine weeks last year.

The saw-tooth is the promotion cycle. Every second week is roughly EUR 8m bigger than its neighbour, which is why getting the cycle right matters more than the model.



Median of the corrected rule, Fourier regression, Theta and the daily gradient boosting model, taken shop by shop, then summed.

The forecast range, and what it assumes

A single number with no range is a guess wearing a suit. The range below is not invented: it is built from how this exact method actually missed across its own 49 test weeks.

August and September 2015, chain total	EUR m	vs last year
the forecast, promotion plan as assumed	327.0	+2.6%
the 80% range, from 49 weeks of measured misses	312.3 to 345.0	−2.0% to +8.2%
if the promotion cycle flips against the plan	317.4	−0.5%
the same nine weeks last year, actual	318.8	

The one assumption that matters: the forecast assumes the fortnightly promotion cycle continues. A real company does not assume this, it reads its own promotion plan. If the cycle lands the other way the answer moves by **EUR 9.6m, or 2.9%**. That is bigger than the gap between the best method and the third best, which is the point.

Public and school holidays and opening days for the forecast weeks are taken from the same week one year earlier, per shop. Stated here rather than buried.

↪ So what

What a finance team should do

Do: fix the comparison week first

Before buying anything, check whether "the same week last year" was the same kind of week. Same promotion, same holidays, same trading days. On this data that one change is worth 6 points of 7.

Do: score the total, not just the parts

Pick the method on the number you actually report. A method that wins shop by shop can lose on the chain total, and the budget is a chain total.

Do: publish a range with every number

Built from the method's own past misses, not from a formula. On this chain that is about ±5% on a single week. A range that has been checked is worth more than a point estimate that has not.

Do not: start with a model

The machine learning is worth having, but it arrived fourth. The corrected spreadsheet rule got to 4.16% on its own, and a finance team can read it, argue with it and sign it.

Method, every measured number and every mistake made along the way are written up in full in the study notes.