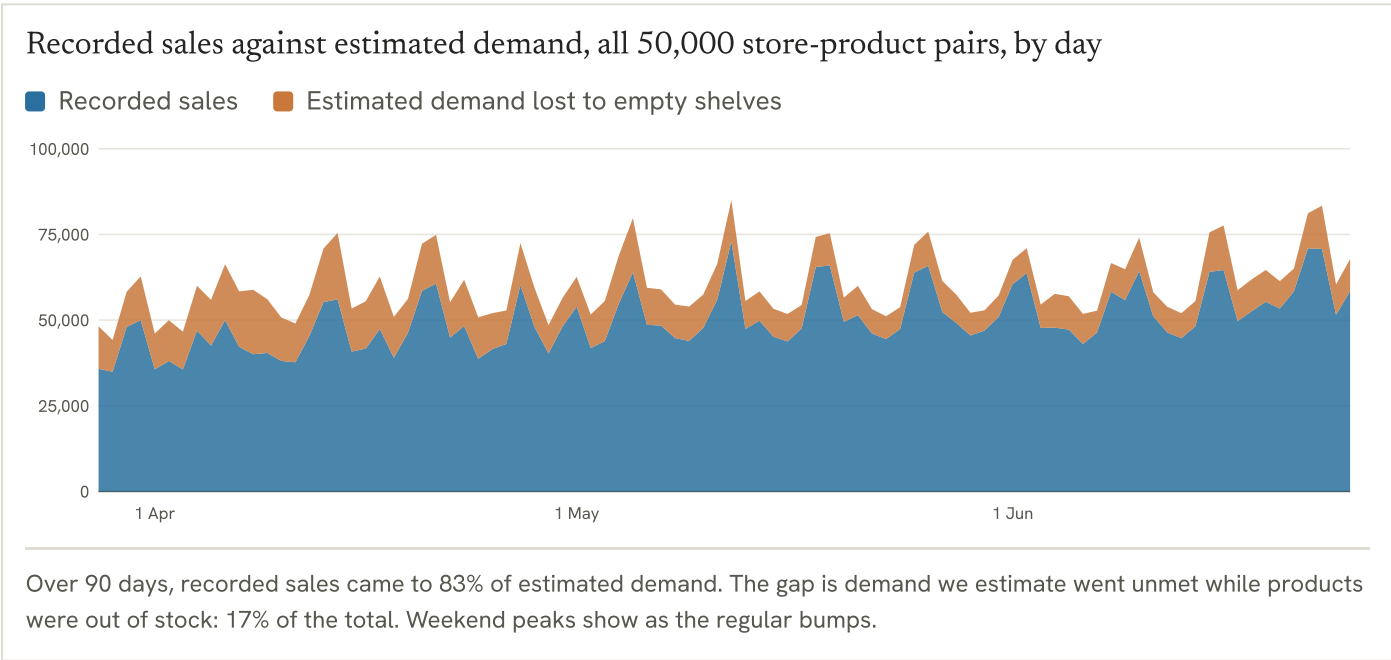


About one sale in six was lost to an empty shelf. Planning from the forecast range cut shortage and waste costs by 45%.

50,000 store-product pairs across 898 stores of an online fresh-grocery retailer, with hourly sales and an hourly out-of-stock flag. We estimated what the empty hours cost, built a seven-day forecast with a range, tested it against simple rules, and turned it into a daily stock plan.

Independent sample using public data, to show what a Dunwell Stock & Demand Planner delivers. Sales are normalised by the publisher, so nothing here is in units or money. Every figure on this page comes from one results file.

Business: online fresh grocery (Dingdong), 18 cities in China History: 28 Mar 2024 to 25 Jun 2024
Test week: 26 Jun to 2 Jul 2024 Prepared by Dunwell Insights, October 2026



Shelves ran empty every evening.
A forecast with a range, used
the right way, fixed most of it.

17%	77%	11%	45%
Of demand lost to empty shelves estimated, 90 days of history	Of zero-sale days had an empty shelf not low demand	Less forecast error than a 7-day average test week, 32.0% against 36.1%	Lower shortage and waste cost test week, central case, against the best simple rule

What we found

- 1

Empty shelves are an evening problem, and every store has it.

44% of product-days had at least one empty hour between 06:00 and 22:00. Products ran out as the day went on: 5% were out of stock at the quietest morning hour, 41% by 21:00. Eight in ten stores had a gap on between 41% and 49% of product-days, so this is about how stock is planned, not a few badly run stores.
- 2

Most zero-sale days were empty shelves, not low demand.

77% of the days a product sold nothing had an out-of-stock hour. Treating those zeros as "nobody wanted it" would teach any forecast to order even less.
- 3

We estimate 17% of demand went unmet.

On days with an empty hour, estimated demand was 1.44 times recorded sales. The common shortcut of scaling sales by hours in stock got 86% of the hidden sales wrong in our test, almost as bad as ignoring them. A model using each product's usual hourly pattern got that down to 43%, with no overall bias. The top 10% of pairs hold 43% of the lost demand.
- 4

The forecast beat both simple rules every week.

Error fell from 36.1% for a 7-day average to 32.0% on the test week, and by a similar margin in each of the four back-test weeks. Knowing the real weather in advance made no difference (32.0%), so a weather feed isn't worth paying for here.
- 5

The textbook stock rule fails when stock keeps for more than a day.

Ordering up to the "cost ratio" point of the forecast assumes anything unsold is thrown away that night. With two days of shelf life it left 7.4 units short per 100 of demand and cost more than the simple rule (15.7 against 13.9). Tuning the stock level on the back-test weeks fixed it.
- 6

Same amount of stock, better placed.

In the central case the tuned plan and the best simple rule ordered almost the same total (115 and 114 per 100 of demand). The plan put it where demand was: shortages fell from 3.1 to 0.6 and waste from 7.6 to 6.5. It won in all 15 of 15 cost and shelf-life scenarios, and on the 5,874 pairs that never ran out that week (12.5 against 17.4), where no demand estimate is needed.

What we'd do

- Plan each day's stock from the forecast range, not from last week's average. Use the stock level tuned for your own shortage and waste costs and shelf life. In the central case (shortage costs twice what waste costs, two days of shelf life) that cut total cost from 13.9 to 7.6 per 100 units of demand on the test week. Resampling whole stores, the saving ranged from 6.0 to 6.5 per 100.
- Stop treating empty-shelf days as low demand. Feed the forecast with the out-of-stock flag, and review the products in the availability list first. They run out early and often.
- Put real costs in before acting. The cost ratio and shelf life here are assumptions. The dashboard shows how the plan changes when you set your own.

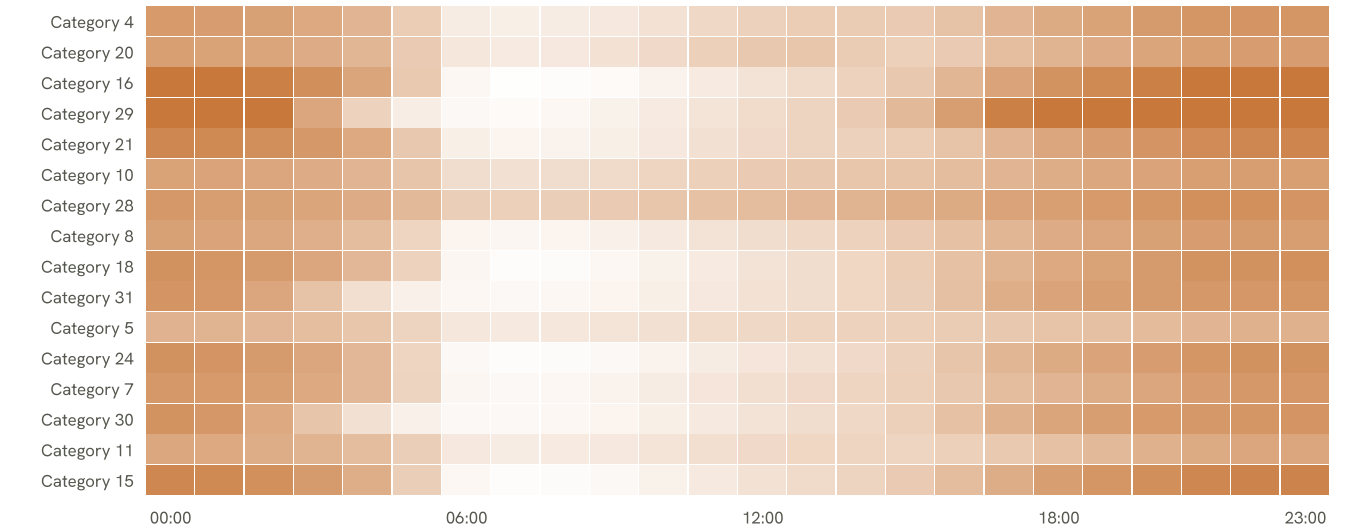
1 · EMPTY SHELVES

44% of product-days had an empty hour, mostly in the evening

The data has an out-of-stock flag for every product and hour. 20% of all open hours (06:00 to 22:00) were flagged. The flag is a reliable sign the shelf was empty: products sold something in 29% of in-stock hours but only 3% of flagged hours, usually the hour they ran out. Every one of the 50,000 pairs ran out at least once, so the publisher may have picked products with stock problems.

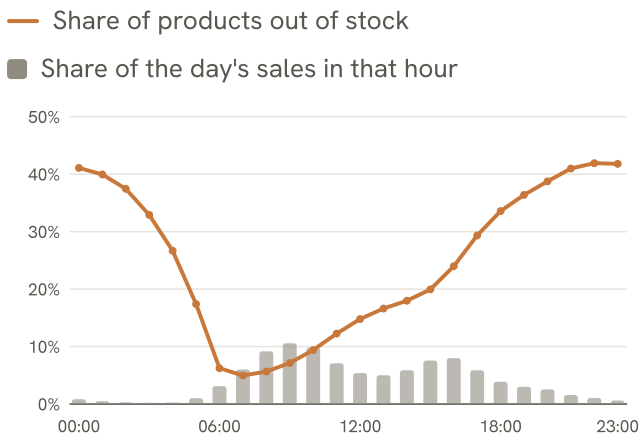
Share of products out of stock, by hour of the day

History, 28 March to 25 June 2024. Darker means emptier. Categories are numbered by the publisher; their names are not given.



The pattern is the same in every category: well stocked in the morning after the overnight delivery, then emptying through the afternoon and evening, when people order for dinner. Night hours look empty because many products are not restocked overnight.

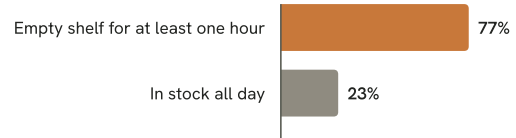
Shelves empty as the day goes on



Sales peak at 09:00 to 10:00, with a second, smaller peak at 16:00 to 17:00. By then about 24% of products are already out, and the share keeps rising.

Why a product sold nothing

Days with zero recorded sales, history



4% of product-days had no sales at all. Only 23% of those were in stock all day.

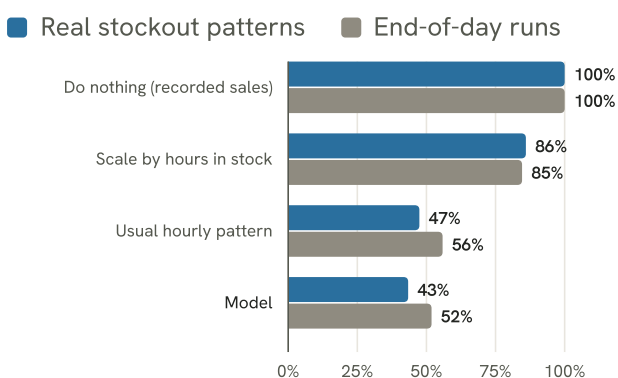
2 · HIDDEN DEMAND

We estimate 17% of demand went unmet, and the usual shortcut gets it badly wrong

Nobody can see what would have sold in an empty hour, so we tested the methods on hours where we know the answer. We took 895,434 days that were fully in stock, hid some hours using real stockout patterns copied from other days, and asked each method to rebuild them. The model learned on half the stores and was scored on the other half.

How well each method rebuilt hours we hid

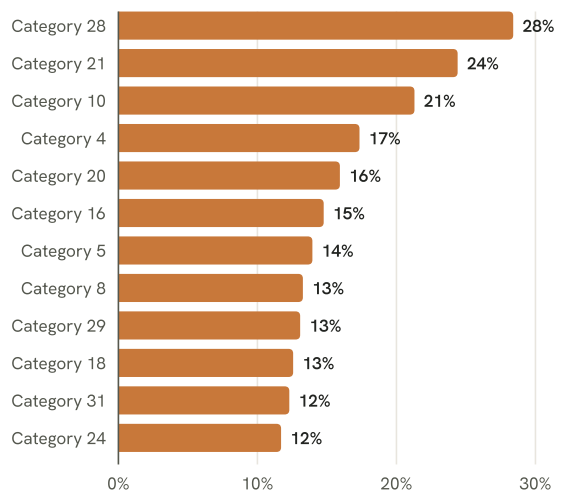
Error as a share of the hidden sales. Scored on stores the model never saw. Lower is better.



Scaling up by hours in stock assumes every hour sells the same, but evenings sell less than mornings. It missed by 86% and was 15% too low overall on real patterns. The model was 9% better than the hourly pattern on its own (range 8% to 9%), so we used it.

Demand lost to empty shelves, by category

Estimated lost demand as a share of estimated demand, history



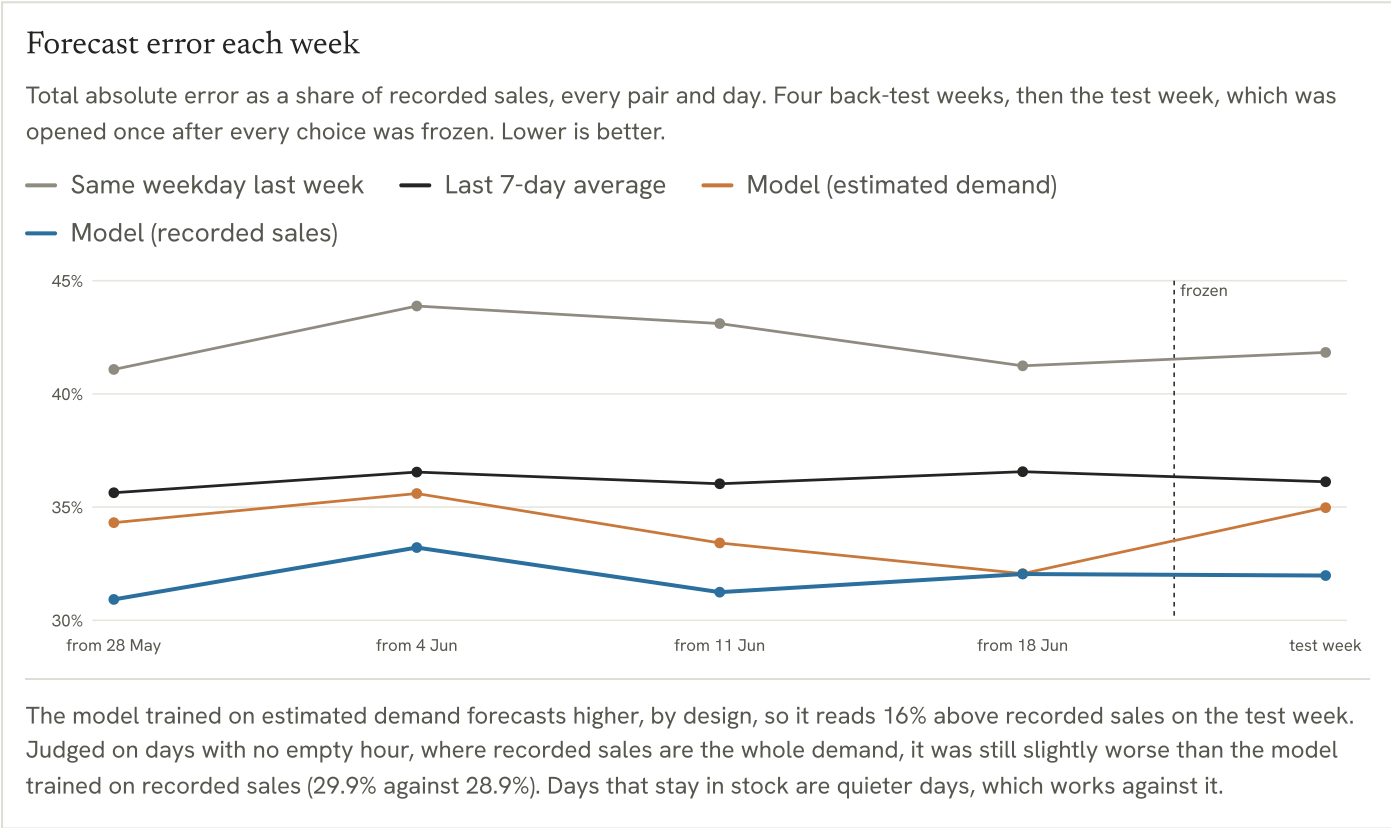
The 12 biggest categories by sales. Every category loses demand. The worst lose over a fifth of it. Half of all pairs lose between 11% and 18%.

A limit to keep in mind: days that stay in stock are quieter than days that run out, and shelves empty because demand was high. Our test can't fully copy that, so the true loss could be higher than 17%. Every estimate here is labelled as estimated demand and kept apart from recorded sales.

3 · FORECASTING THE NEXT SEVEN DAYS

The model cut forecast error by 11% against a 7-day average, and held up on the test week

Each forecast is made at the end of a day for the next seven. We compared two simple rules with a LightGBM model trained on all pairs at once. It uses recent sales, the same weekday in past weeks, how often the product ran out, the planned discount and promotion, the category, store size and the day of the week. Everything was fitted only on data before each cutoff. We picked every setting on four back-test weeks, froze them on 11 Oct 2026 at 16:40 UTC, then scored the test week once.

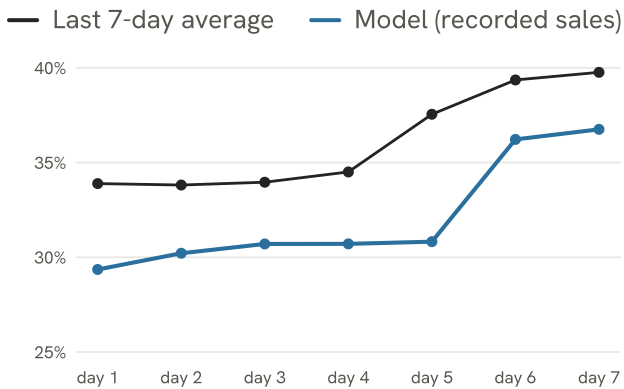


Every forecast on the test week						
Forecast	Error, all days	Error, in-stock days	Bias	Scaled error	Quantile loss	80% range caught
Same weekday last week	41.8%	38.9%	+0.3%	1.12	-	-
	back-test 42.3%					
Last 7-day average	36.1%	32.5%	+0.3%	0.96	0.134	81%
	back-test 36.2%					
Model, trained on recorded sales	32.0%	28.9%	+1.7%	0.85	0.110	75%
	back-test 31.9%					
Model, trained on estimated demand	35.0%	29.9%	+15.9%	0.93	0.127	70%
	back-test 33.8%					
Model with the real weather (check only)	32.0%	29.0%	-2.4%	0.85	-	-
	back-test 31.9%					

Error is total absolute error as a share of recorded sales. Scaled error compares each pair with its own week-to-week change; under 1 beats "same weekday last week". Quantile loss scores the whole range. The model's 80% range is 35% narrower than the simple rule's but caught 75% of days, not 80%. The tuned stock level corrects for that. Quantiles crossed on 11% of forecasts before sorting.

Error by how far ahead

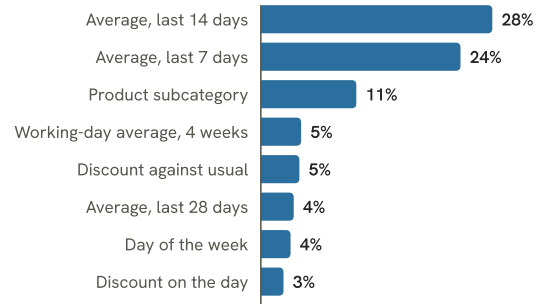
Test week, days 1 to 7 after the cutoff



Error grows the further ahead you look, and jumps for the last two days, a Monday and Tuesday at the start of July. The model stays ahead at every step.

What the model leans on

Share of the model's total gain from each input, test-week model trained on recorded sales



Recent averages carry most of the weight. The product subcategory and the planned discount come next. Weather at the cutoff barely registers.

Where the model helps most

Test-week error by sales level and by how often the pair ran out in the four weeks before

Group	Same weekday	7-day average	Model	Model gain
Sales level: low 0.04 to 0.57 a day	53.8%	45.1%	41.2%	9%
Sales level: middle 0.57 to 0.93 a day	49.1%	41.3%	36.2%	12%
Sales level: high 0.93 to 30.95 a day	36.6%	32.3%	28.4%	12%
Ran out: low 0% to 29% of days	45.3%	38.8%	34.0%	12%
Ran out: middle 32% to 43% of days	42.9%	37.2%	33.0%	11%
Ran out: high 46% to 100% of days	39.4%	34.1%	30.3%	11%

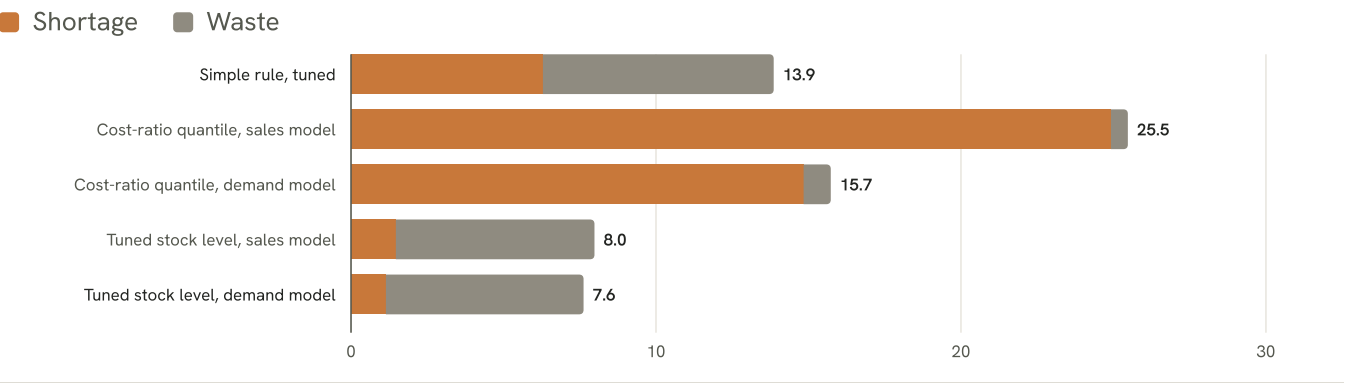
Relative error isn't defined for pairs that sold nothing in a period. We report total error over total sales instead, which stays defined. 59% of test-week pair-days had no empty hour. Ranges for the gap between model and rule resample whole stores, so related products and days stay together: on the test week the model's error was 4.1 points lower (range 4.0 to 4.3).

With the stock level tuned, the forecast plan cost 45% less than the best simple rule

We ran a week of daily deliveries for every pair. Each morning the shelf is topped up to a target. Customers buy the oldest stock first. Stock older than its shelf life is thrown away at night. Shortage and waste costs are assumptions, not data, so we show five cost ratios and three shelf lives. The simple rule orders a multiple of the last 7 days' average. We tuned that multiple on the back-test weeks so it gets its best shot (1.95 times in the central case). The model plan got one tuned setting too.

Test week, central case: 26 Jun to 2 Jul

Cost per 100 units of demand, in relative units. Shortage costs twice what waste costs, stock keeps two days. A week of daily deliveries, every pair.



The plain cost-ratio rule leaves too many shortages because it treats unsold stock as wasted the same night. Tuned on the back-test weeks, the model plan lands at 7.6, against 13.9 for the tuned simple rule. The recommended plan uses the forecast trained on estimated demand, which had the lower cost on the back-test weeks. The one trained on recorded sales came close (8.0), so most of the gain comes from the forecast and the tuning, not from the demand estimates.

Every scenario on the test week

Cost per 100 units of demand: tuned model plan against the tuned simple rule. The change is the model plan's cost minus the rule's.

Shortage : waste cost	Shelf life	Simple rule	Model plan	Change	Demand met
1 : 3	1 day	47.9	40.8	-15%	76% rule 69%
1 : 3	2 days	14.8	9.1	-38%	96% rule 93%
1 : 3	3 days	5.8	3.3	-42%	99% rule 97%
1 : 1	1 day	32.4	27.3	-16%	89% rule 82%
1 : 1	2 days	9.2	5.6	-40%	99% rule 96%
1 : 1	3 days	3.7	1.9	-49%	100% rule 98%
2 : 1	1 day	48.5	37.9	-22%	93% rule 88%

Shortage : waste cost	Shelf life	Simple rule	Model plan	Change	Demand met
2 : 1 (central)	2 days	13.9	7.6	-45%	99% rule 97%
2 : 1	3 days	5.7	2.2	-62%	100% rule 98%
4 : 1	1 day	69.3	50.6	-27%	97% rule 93%
4 : 1	2 days	20.7	8.8	-58%	99% rule 98%
4 : 1	3 days	8.9	2.9	-68%	100% rule 99%
9 : 1	1 day	99.5	67.1	-33%	99% rule 96%
9 : 1	2 days	32.6	11.8	-64%	100% rule 98%
9 : 1	3 days	15.4	4.0	-74%	100% rule 99%

The model plan won in all 15 scenarios. The gain is smallest when stock keeps only one day, because there is no carry-over to manage, and largest when shortages are expensive and stock keeps longer.

Demand in these scenarios is recorded sales plus our estimate for the empty hours, 14% above recorded sales on the test week. As a check we reran the central case on the 5,874 pairs that never ran out that week, where no estimate is needed. The plan still cost less: 12.5 against 17.4. The week starts with an empty shelf, and stock still fresh at the end is not counted.

5 · NEXT STEPS

Plan from the range, fix the evening gaps, and measure real costs

- 1 Put your real costs in.**
What does a lost sale cost you, and what does a thrown-away unit cost? How many days does each product keep? The stock level depends on all three.
- 2 Start with the availability list.**
It ranks every pair by estimated lost demand. The top of the list runs out most days, often by mid-afternoon.
- 3 Run the plan beside your current ordering for four weeks.**
Keep some stores on the old way so the difference can be measured, the same way we did here.
- 4 Record what was on the shelf, not just what sold.**
A daily stock count, or even a note of when each product ran out, makes demand far easier to estimate than an hourly flag.
- 5 Check the range each month.**
The model's 80% range caught 75% of days. If that drifts, retune the stock level.

The plan was written down before any forecast was scored

The analysis plan was locked before any forecast was fitted. One change was made later, before the test week was forecast, and it is dated in the plan: the tuned model plan was added after the first back-test week showed the textbook rule failing. Every choice was then frozen with a fingerprint (a0cf165ecee6...) before the test week was scored. The whole study reruns from the two raw files with one command.

1 Check the data One row per pair and day, hourly sales add up, the flag meaning confirmed	2 Rebuild hidden demand Hide hours on in-stock days and score four methods	3 Forecast and back-test Four weeks, each fitted only on data before its cutoff	4 Freeze, then test Score the test week once and run the stock scenarios
--	--	---	--

The methods <table><tr><td>Demand recovery</td><td>None, scaling by hours in stock, usual hourly pattern, and a LightGBM model on top of the pattern</td></tr><tr><td>Forecasts</td><td>Same weekday last week, 7-day average, LightGBM quantile models (5% to 95%) on recorded sales and on estimated demand</td></tr><tr><td>Check only</td><td>The same model given the real weather on the day</td></tr><tr><td>Stock plans</td><td>Simple rule (multiple of the 7-day average), cost-ratio quantile, and tuned stock level</td></tr></table>	Demand recovery	None, scaling by hours in stock, usual hourly pattern, and a LightGBM model on top of the pattern	Forecasts	Same weekday last week, 7-day average, LightGBM quantile models (5% to 95%) on recorded sales and on estimated demand	Check only	The same model given the real weather on the day	Stock plans	Simple rule (multiple of the 7-day average), cost-ratio quantile, and tuned stock level	Checks we ran • 50,000 pairs × 97 days, no gaps or duplicates • Hourly sales add up to the daily total (largest gap under 0.00001) • Flag 1 means empty: sales in 2.9% of flagged hours against 28.8% of others • Recovery model scored only on stores it never saw • Each back-test week fitted on data before its cutoff, recovery included • Test-week split checked in the data: 26 Jun to 2 Jul • Choices frozen before the test week; scoring refuses to run without the freeze file • Stock result rechecked on pairs that never ran out
Demand recovery	None, scaling by hours in stock, usual hourly pattern, and a LightGBM model on top of the pattern								
Forecasts	Same weekday last week, 7-day average, LightGBM quantile models (5% to 95%) on recorded sales and on estimated demand								
Check only	The same model given the real weather on the day								
Stock plans	Simple rule (multiple of the 7-day average), cost-ratio quantile, and tuned stock level								

Data dictionary

Field	What it is
store_id, product_id	Store and product codes. One pair is one product in one store.
city_id, category ids	City and three levels of product category, as numbers. Names are not published.
sale_amount	Daily sales, normalised by an unpublished factor. Not units or money.
hours_sale	24 hourly sales figures that add up to the daily total.
hours_stock_status	24 hourly flags. 1 means out of stock in that hour.
discount	Price paid as a share of full price. 1 means no discount.

Field	What it is
activity_flag, holiday_flag	Promotion running; weekend or public holiday.
precpt, avg_temperature, avg_humidity, avg_wind_level	Daily weather for the store.

Decisions log

Decision	Why
Use all 50,000 pairs	No picking of easy products. A fixed sample of 5,000 was used only to write the code.
Keep recorded sales and estimated demand apart	Estimates are never used as the truth when scoring forecasts.
Score forecasts on days with no empty hour as well	There, recorded sales are the whole demand.
Leave out future weather	A shop doesn't know it when it orders. Shown once as a check.
Allow planned discounts and promotions	The shop sets these in advance.
Tune the simple rule as well as the model	So the model has to beat the rule at its best.
Resample whole stores for ranges	Products in one store share deliveries and customers.

What this data can't tell us

- Sales are normalised by an unpublished factor. They are not units, kilos or money, so no figure here is in pounds.
- Demand during an empty hour is never seen. Our estimates were tested by hiding hours on in-stock days, which are quieter than the days shelves really ran out. The real loss could be higher or lower.
- The out-of-stock flag shows when a product was unavailable. It is not a stock count, so we can't see how much was on the shelf or delivered.
- The stock scenarios use assumed costs, a daily delivery, an empty shelf at the start of the week and a fixed shelf life. Real lead times, minimum orders and pack sizes would change the numbers.
- The demand in the scenarios includes our estimates for empty hours. We checked the result on pairs that never ran out in the test week, but those are not typical pairs.
- The data covers 14 weeks of fresh grocery in China in spring and summer 2024. It says nothing about yearly patterns, and anything about fashion or long-life products needs its own evidence.

This analysis is based on the data described here and the methods shown. It supports decisions. It is not a promise of any result, and it is not financial, legal or tax advice.

Every job. Done well.

Dunwell Insights · info@dunwellstudio.com

Data: FreshRetailNet-50K, Dingdong Inc. (2025), CC BY 4.0. Wang et al., "FreshRetailNet-50K", arXiv 2505.16319. huggingface.co/datasets/Dingdong-Inc/FreshRetailNet-50K. We reshaped the data and added estimates; the published figures are normalised. Dingdong does not endorse Dunwell, and this is not a Dunwell client.