

Can better models boost the dashboard figures?

Before and after figures for every model on binzagr.snop.live, how each "after" was reached, and whether it is more accurate.

Version 1.0 • 23 October 2026 • Site read on 23 Oct 2026: Value cockpit to Value roll-up (14 pages)

Data: 02_Cleaned v3, the same files the site uses. Our hold-out sales match the site to the cent.

Companion data: **06_Model_Review/data** (before_after_figures.csv / .json, every test table, improved picks and churn scores)

+0.025 churn AUC on unseen outlets (0.772 to 0.797), picked before the test	+0.7 pts next-best SKU picks bought (11.6% to 12.4%) on the site's own test	-1.4 pts SKU forecast error at the same bias (66.9% to 65.5%)	+SAR 12K a year on the SAR 3.45M base case (under 1%)
---	---	---	---

1. The answer

Better models lift three figures a little and leave the rest where they are. The forecast ensemble is already close to the best this data allows, and some headline figures go down when they are measured fairly. The limit is the data: eight months of history and four test weeks, not the choice of algorithm.

- **Churn warning: real boost.** Extra trees, picked on validation, scores AUC 0.797 against 0.772 for LightGBM (the site's model type) on the same unseen outlets: +0.026, 90% interval +0.011 to +0.040. Its riskiest 20% holds 59% of the outlets that went quiet, against 52%.
- **Next-best SKU: small boost.** On the site's own test (picks on 25 Aug, 4,097 outlets, 8 weeks) a LightGBM ranker over new scores (EASE, RP3beta) raises the share of picks bought from 11.6% (our copy of the site scorer; the site reports 12.0%) to 12.4%, 90% interval +0.5 to +0.9 points. The top 5, which the value uses, barely moves.
- **Cluster adoption: boost in our copy, but the site figure needs checking.** Blending logistic regression with XGBoost lifts the within-cluster AUC from 0.71 to 0.75. With a time-honest split no model reproduces the site's 0.90 (we reach 0.78 to 0.79).
- **Demand forecast: keep it, show the bias.** At SKU level one equal blend of the best three models is 1.4 points better than the site at the same bias. At van, principal and cluster level the site's blends are 0.6 to 1.4 points better than anything we could build without its foundation models.
- **Goes down when measured fairly:** the "22% less SKU error" falls to about 15% if the forecast must be unbiased; the principal-level ML loses to a 4-week weekday average on the page as shown; the company-total comparison cannot rank models at all in four weeks.
- **Value:** model accuracy drives only about SAR 217K of the SAR 3.45M. All upgrades together add about SAR 12K a year. The large levers (fleet SAR 1.39M, labour SAR 639K, price SAR 397K) rest on benchmarks and assumptions that better data, not better models, would firm up.

How to read "before" and "after". Before = the figure on the site. After = the same test re-run on the same data with the best model or combination, chosen on earlier validation windows only; the test weeks were used once. Where the site's exact model could not be run (its three foundation models), we compare against the site's own published result and show our copy of its method beside it.

2. Before and after: every figure to change on the site

One row per figure, page by page, in the words the site uses. "On the site now" is the text to find; "Replace with" is the text to put in its place. The same table is in data/before_after_figures.csv and .json, with the numbers in separate columns and the basis of every after value.

Type	What to do		
Replace	Measured on the same test after switching the model; change it once the new model is live on the site		
Fix	A wrong or mislabelled figure; change it now, no model change needed		
Add	Keep the figure and add the note		
Verify first	Our time-honest re-test is lower; re-run the site test on a clean split, then use whichever result it gives		
Recompute	Needs the site pipeline re-run; we cannot produce the exact figure		
Keep	No change		

#	Page and place	On the site now	Replace with	Type
1	Value cockpit KPI tile, Forecast proof	22% · less error than last week repeated	22% · less error than last week repeated (forecast 15% below actual)	Add
2	Value cockpit Proof card: Demand forecast	ML forecast 64% · 4-week weekday average 72% · Repeat last week 82%	Keep the three values; add "ML bias -15%" under ML forecast	Add
3	Value cockpit Proof card: Next-best SKU	3.0× · Model picks 12% · Best-sellers 4.0% · Random SKUs 0.7%	3.3× · Model picks 12.4% · Best-sellers 3.8% · Random SKUs 0.8%	Replace
4	Value cockpit Proof card: Outlet churn warning	0.77 · Churn model 0.77 · Recency rule 0.68 · top 20% of risk holds 48% of churners	0.80 · Churn model 0.80 · Recency rule 0.73 · top 20% of risk holds 59% of churners	Replace
5	Value cockpit Proof card: Cluster adoption	65% · Adoption model 65% · Channel popularity 40% · What neighbours buy 12% · AUC 0.90	32% · Adoption model 32% · Channel popularity 30% · What neighbours buy 9% · AUC 0.79	Verify first
6	Value cockpit Scorecard: Next-best SKU	vs best-sellers · 3.0× as good	vs best-sellers · 3.3× as good	Replace
7	Value cockpit Scorecard: Outlet churn warning	vs recency rule · +0.09 AUC	vs recency rule · +0.07 AUC	Replace
8	Value cockpit Scorecard: Cluster adoption	vs channel popularity · 1.6× as good	vs channel popularity · 1.1× as good	Verify first
9	Value cockpit Scorecard: Demand forecast	vs 4-week weekday average · 12% less error	Keep; add "3% at zero bias"	Add
10	Value cockpit Recommendation 1 (leads)	AUC 0.902 on unseen outlets	AUC 0.79 on a time-honest test (0.75 within cluster)	Verify first
11	Value cockpit Recommendation 3 (P1 accounts)	SAR 927,163 past due (24%)	SAR 590,932 past due (15%) + SAR 336,231 older than the invoice window	Fix
12	Value cockpit Recommendation 5 (win-back)	churn model AUC 0.77 on a later period it never saw · Top 20% risk holds 48% of churners	churn model AUC 0.80 on outlets it never saw · Top 20% risk holds 59% of churners	Replace
13	Value cockpit Explore card: Receivables	SAR 927K past due, 24% of the credit book	SAR 591K past due, 15% of the credit book (+ SAR 336K unmatched)	Fix
14	Value cockpit Explore card: Next-best SKU	12% of the model's picks were bought · 3.0× best-sellers	12.4% of the model's picks were bought · 3.3× best-sellers	Replace
15	Value cockpit Hero and impact tree	SAR 3.45M a year	Keep (all model upgrades add about SAR 12K, under 1%)	Keep
16	Forecast proof Hero	the ML forecast made 22% less error than repeating last week	Keep; add "with the forecast 15% below actual sales"	Add
17	Forecast proof Tile: Van-week error removed	38.6% vs 47.2% weekly error	40.0% vs 47.2% weekly error	Fix
18	Forecast proof Section heading (Model vs planner)	Less error at every level where orders are placed	Less error at SKU and van level; level with a simple average above that	Fix
19	Forecast proof Principal level chip	-8% error	Level with the 4-week average (28.3% vs 27.4%)	Fix
20	Forecast proof Company total level	About level · ML 14.6% vs 14.9%	Keep; add "four weeks cannot rank models at total"	Add
21	Forecast proof SKU level: ML forecast	63.8%	Keep until re-run; expect about 1.4 points lower at the same bias	Recompute

#	Page and place	On the site now	Replace with	Type
22	Forecast proof Van, principal, cluster: ML forecast	40.0% · 28.3% · 21.7%	Keep	Keep
23	Forecast proof Model race text	Its forecast is a blend of the best one to three models	SKUs: one equal blend of the three models best on validation. Vans: a blend of the best one to three	Recompute
24	Forecast proof Next 12 weeks: weekly range	The weekly range is the sum of daily p10 and p90	Weekly bands calibrated on validation errors	Recompute
25	Forecast proof Cluster cards: days in range	Ice Cream 54% · Tea & Dairy Foods 93% of days in range	Bands calibrated per cluster (target 80%)	Recompute
26	Forecast proof Next 12 weeks: company model	Company model: LightGBM + AutoETS + TiRex	Company model: TiRex + AutoETS + AutoTheta (the blend tested on the hold-out)	Fix
27	Next-best SKU Hero	outlets bought 12.0% of their ten recommended SKUs, against 4.0% for the best-sellers and 0.7% for random SKUs	outlets bought 12.4% of their ten recommended SKUs, against 3.8% for the best-sellers and 0.8% for random SKUs	Replace
28	Next-best SKU Tile	58.2% bought at least one pick	61.4% bought at least one pick	Replace
29	Next-best SKU Hold-out proof	4,097 outlets judged. Lift: 3.0× over best-sellers, 18.2× over random.	4,097 outlets judged. Lift: 3.3× over best-sellers, 16.3× over random.	Replace
30	Next-best SKU New line under the proof	(not shown)	Top five picks: 14.6% bought (the list the value uses)	Add
31	Next-best SKU Method text	similar outlets' baskets, neighbourhood and van penetration, popularity	similar outlets' baskets, neighbourhood and van penetration, popularity, plus EASE and RP3beta scores, ranked by a LightGBM ranker	Replace
32	Outlets Hero	AUC 0.77 against 0.68 for the recency rule	AUC 0.80 against 0.73 for the recency rule	Replace
33	Outlets Tile: riskiest 20%	47.9% of the outlets that lapsed sat in the top 20% of risk	59.2% of the outlets that lapsed sat in the top 20% of risk	Replace
34	Outlets Text under the AUC	Top risk decile churns at 2.9× the average rate (6.7% in the test window)	Top risk decile churns at 4.0× the average rate (7.8% in the test window)	Replace
35	Outlets Test description	trained to 21 Aug	trained on cut-offs to 30 Jun, tested at 25 Aug on outlets it never saw	Replace
36	Outlets Risk tiers	High risk 57 outlets · Medium risk 129 outlets	Re-score every outlet with the new model (data/churn_scores_v2.csv)	Recompute
37	Cluster harmonisation Tile	0.90 model AUC on unseen outlets	0.79 model AUC on unseen outlets (time-honest test)	Verify first
38	Cluster harmonisation Proof: top 10% of ranked pairs	Adoption model 65% · Channel popularity 40% · Neighbours buy it 12%	Adoption model 32% · Channel popularity 30% · Neighbours buy it 9%	Verify first
39	Cluster harmonisation Proof: lift text	6.5× more adoptions in the model's top 10% than in a random 10% (29.3% of top-10% pairs adopted vs 4.54% on average)	3.2× more adoptions in the model's top 10% than in a random 10% (18.0% of top-10% pairs adopted vs 5.66% on average)	Verify first
40	Cluster harmonisation Ranking accuracy (AUC)	Adoption model 0.90 · Channel popularity 0.71 · Neighbours buy it 0.56	Adoption model 0.79 · Channel popularity 0.73 · Neighbours buy it 0.46	Verify first
41	Cluster harmonisation How well it ranks each cluster	Per-cluster AUC 0.67 to 0.86	Keep; add "mean within cluster 0.75" next to the pooled AUC	Add
42	Receivables Headline	SAR 927K past due (24%)	SAR 591K past due (15%) + SAR 336K older than the invoice window	Fix
43	Value roll-up Labour lever evidence	625 badge-sharing sessions flagged	625 high or medium-risk badge sessions (198 high)	Fix
44	Value roll-up ML forecast to safety stock	SAR 65.2K a year · + SAR 296K once	Keep	Keep
45	Value roll-up Whitespace / next-best SKU	SAR 105K a year	Keep	Keep
46	Value roll-up Churn early-warning and win-back	SAR 47.3K a year	Up to SAR 53.4K after the new model is live	Recompute

Replace rows use numbers measured on the same test once the new model is live on the site; until then the current figures stand. Verify-first rows are lower in our time-honest re-test, so they should be re-checked before the demo, not raised.

3. How the after figures were reached

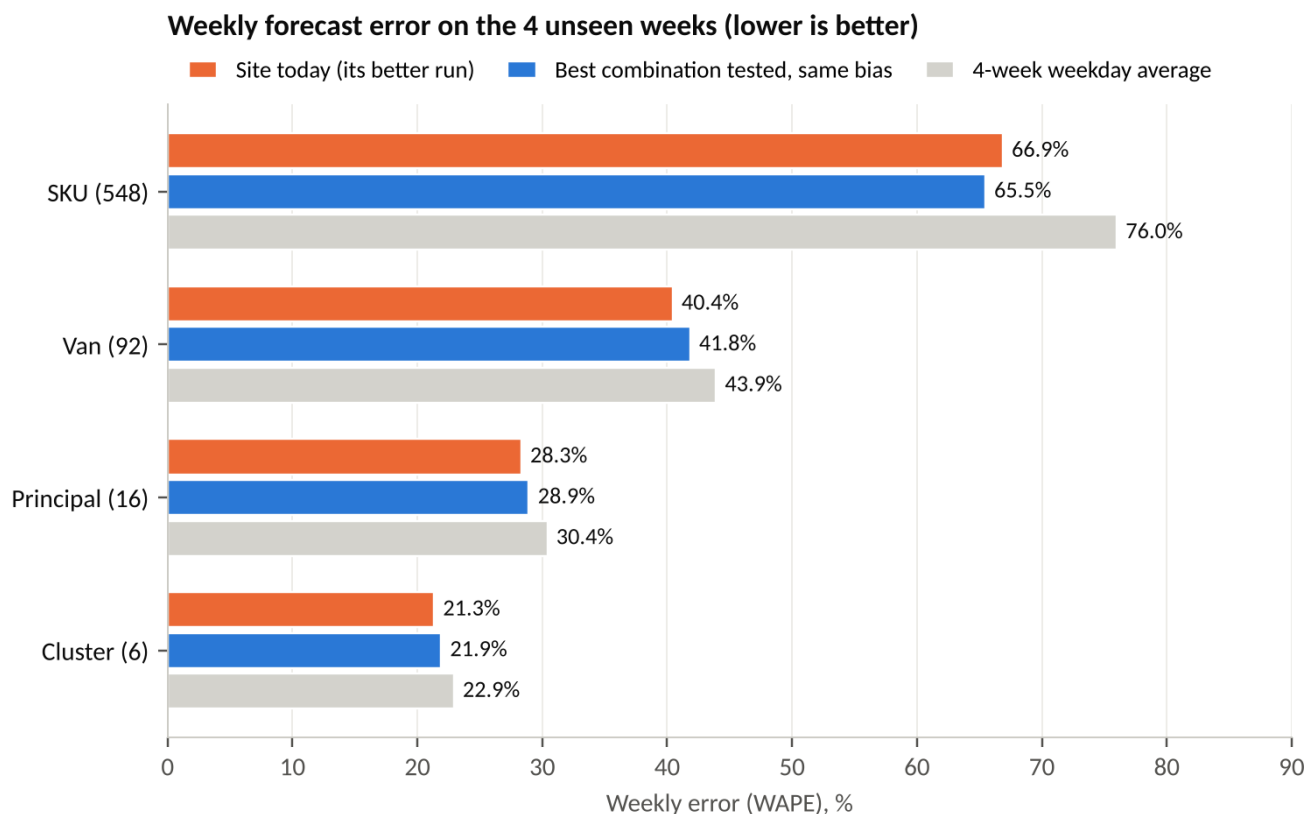
Every test copies the site's own: the same files, the same series, the same test windows and the same comparison baselines. We first checked that our copy reproduces the site's baselines; only then did we try new models.

Check	Site	Our copy
Hold-out van sales, 23 Sep - 20 Oct (company total)	SAR 5,060,755	SAR 5,060,755 (all 92 vans within SAR 0.02)
SKU error: repeat last week / 4-week weekday average	87.5% / 76.0%	87.6% / 76.5%
Van error: repeat last week / 4-week weekday average	53.3% / 43.9%	53.3% / 44.5%
Principal error: repeat last week / 4-week weekday average	35.8% / 30.4%	35.8% / 30.4%
Next-best SKU: outlets judged / best-sellers / random	4,097 / 3.95% / 0.66%	4,097 / 3.8% / 0.8%
Next-best SKU: the site scorer	12.0%, 58.2% of outlets hit	11.6%, 58.5% of outlets hit
Cluster adoption: test adoption rate	4.5% (242 of 5,335 pairs)	5.7% (220 of 3,885 pairs; Personal Care left out)

- **Chosen before the test.** Every model, weight and combination rule was picked on validation windows that end before the test starts (forecast: origins 28 Jul and 25 Aug; next-best SKU: picks at 30 Jun; churn: cut-offs 16 and 30 Jun; adoption: May - Jul). Combination rules were nested: members picked on one window, the rule scored on the next. The test was then run once.
- **Compared fairly.** Forecasts are compared at the same bias, because weekly error rewards forecasting low (section 4). Classifiers are compared on the same held-out outlets, with paired bootstrap intervals for the gain.
- **What we could not run.** The site's foundation models (Chronos-2, TimesFM 2.5, TiRex) cannot be downloaded in our environment. Our forecast pool is LightGBM (daily and weekly direct models) plus seven statistical methods (ETS, Theta, SES, ADIDA, IMAPA, TSB, SBA) on daily and weekly data and the two baselines, so our forecast "after" is a lower bound for what the site could reach by changing its combination rule.

4. Demand forecast (Forecast proof, van loads)

The site blends the best one to three models per series, picked by a backtest race (the ratio is not shown; the race leads are foundation models 31, LightGBM 30 and statistical 31 of 92 vans). We tested ten combination rules on the same 28 days.

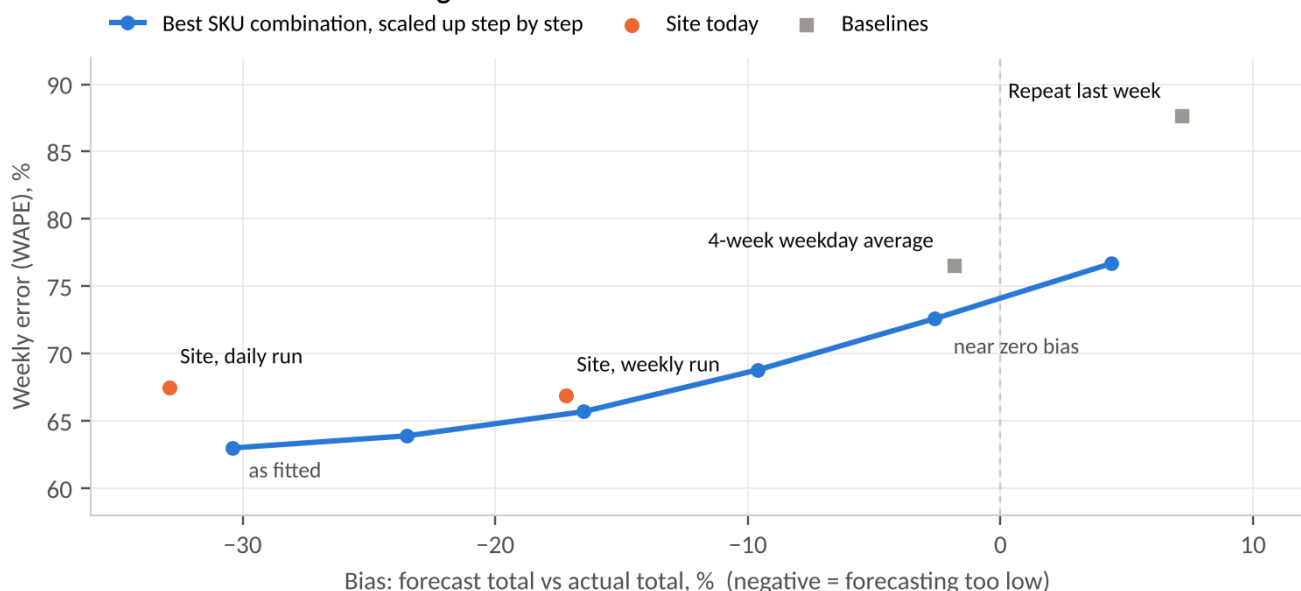


Same data, series, weeks and baselines as the site (raw van sales, 23 Sep - 20 Oct 2026). Company total is left out: four weeks cannot rank models there.

Level	Site today	Our copy of the site style	Best rule (picked on validation)	As fitted	Same bias as site	Verdict
SKU (548)	66.9% at -17% (weekly run)	77.9% at +1%	level top 3 models, equal	63.0% at -30%	65.5%	Small boost
Van (92)	40.4% at -10% (daily run)	45.4% at +1%	median of all models	43.0% at -1%	41.8%	Keep site
Principal (16)	28.3% at -13% (daily run)	33.6% at -2%	mean of all models	30.2% at 0%	28.9%	Keep site
Cluster (6)	21.3% at -7% (daily run)	23.6% at +1%	median of all models	22.5% at +1%	21.9%	Keep site

- **Composition.** At SKU level, racing models per SKU overfits: with two 4-week windows to choose from, the race picks whatever was lucky. Our copy of that rule scored 77.9%, worse than the 4-week average. One blend for all SKUs of the three models that were best on validation (LightGBM weekly + LightGBM daily + IMAPA daily) scored 63.0%.
- **Ratio.** Equal weights. Weights fitted on validation (LightGBM daily 57%, repeat-last-week 17%, Theta 10%, LightGBM weekly 7%, SBA 6%, ETS 3%) scored 63.4% against 63.0% for equal weights: fitted ratios do not beat a plain average with this little data.
- **Van, principal, cluster.** The site's blends, with foundation models, beat everything we could build by 0.6 to 1.4 points at the same bias. Keep them.
- **Company total.** The four test weeks were SAR 1.08M, 1.38M, 1.56M and 1.05M. Any flat forecast between 1.08M and 1.38M a week scores exactly 16.2%, because the misses cancel, so no model can be shown to beat another at total.

SKU level: lower error is bought with a lower forecast

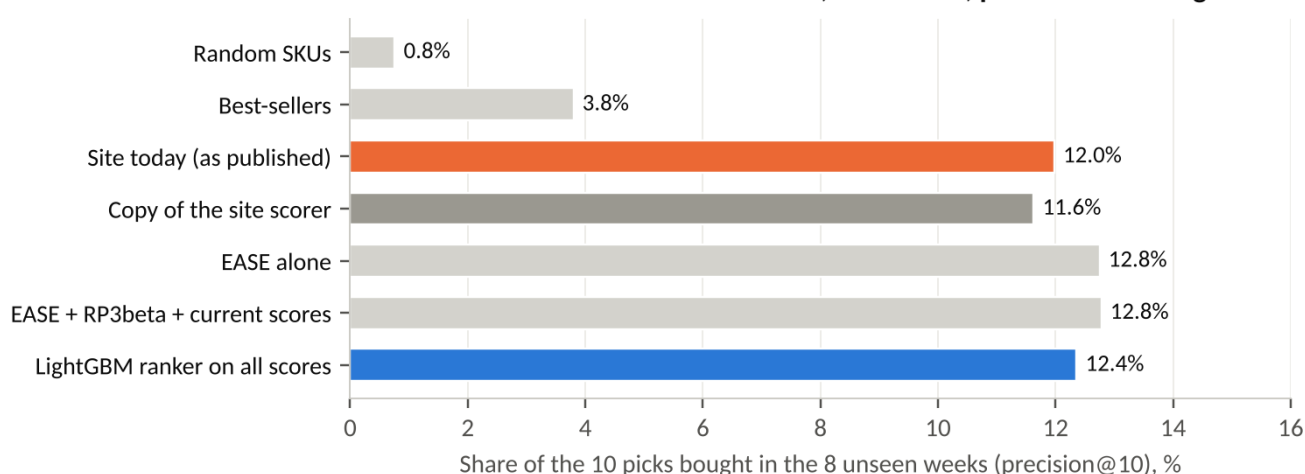


Weekly error (WAPE) is lowest for a forecast below the mean, because SKU sales are lumpy: most weeks are small, a few are large. Both the site's runs and our blend forecast 17% to 33% too low. Held at the site's bias, our blend is 1.4 points better. Forced to zero bias it scores 74.1%: 15% less error than repeating last week (not 22%) and 3% less than the 4-week weekday average. For ordering, a forecast 17% too low means ordering 17% too little unless safety stock covers it, so the page should show the bias next to the error.

5. Next-best SKU

The site ranks SKUs an outlet has not bought by what similar outlets buy, neighbourhood and van penetration, and popularity. We added two stronger recommenders, EASE (a closed-form item model) and RP3beta (a graph walk), and two ways to combine them with the current scores.

Next-best SKU on the site's own test: 4,097 outlets, picks made 25 Aug



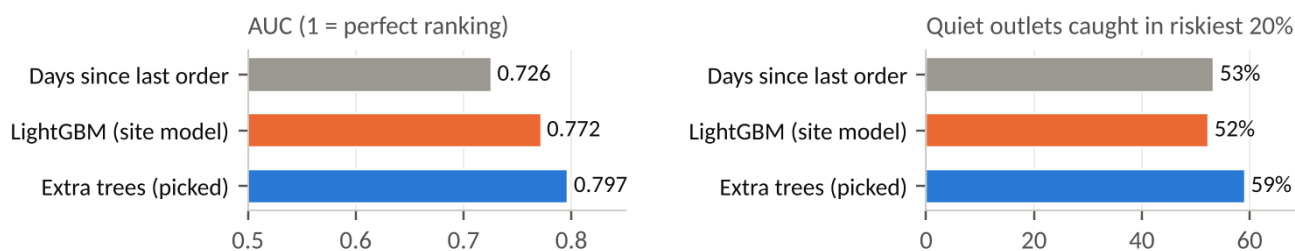
Model	Validation (picks 30 Jun)	Test: share of 10 picks bought	Test: top 5	Test: outlets with a hit
Best-sellers	5.2%	3.8%	3.5%	23.5%
Copy of the site scorer	14.1%	11.6%	14.5%	58.5%
EASE alone	13.9%	12.7%	15.5%	62.4%
RP3beta alone	13.6%	12.6%	15.3%	61.6%
EASE + RP3beta + current scores	14.5% (weights from 2 Jun)	12.8%	15.5%	62.4%
LightGBM ranker (picked)	15.5%	12.4%	14.6%	61.4%

- **Picked fairly.** On validation the ranker scored 15.5% and the blend 14.5% (blend weights tuned on an earlier cut-off, so both are out of sample). The ranker was picked; on the test it scored 12.4%. The blend would have scored 12.8%: the honest choice is not always the best one on the test.
- **Gain.** +0.7 points over our copy of the site scorer, 90% bootstrap interval +0.5 to +0.9. Expect the site's 12.0% to reach about 12.7%.
- **Ceiling.** In the test weeks the 4,097 outlets started buying 5.6 new SKUs each on average, so a perfect list would have 42% of its 10 picks bought. Models reach 12% to 13% because adoption depends on what the rep offers and what is on the van, which the files do not record.

6. Outlet churn warning

Same protocol as our 05_Predictions churn model: will an outlet active in the last 13 weeks make no purchase in the next 8? 30% of outlets are held out; seven model types and every 2- and 3-model blend were compared on two validation folds, then the winner was tested once.

Churn warning on outlets the models never saw (test at 25 Aug)



1,314 held-out outlets, 103 went quiet in the 8 weeks after 25 Aug. Extra trees was picked on two validation folds before the test.

Model	Validation AUC (mean of 2)	Test AUC	Test: quiet outlets in riskiest 20%
Extra trees (picked)	0.912	0.797	59%
Random forest	0.901	0.790	58%

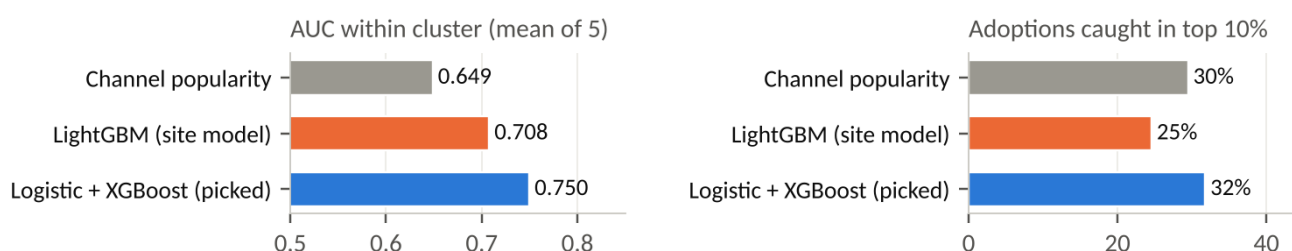
Model	Validation AUC (mean of 2)	Test AUC	Test: quiet outlets in riskiest 20%
Blend: all seven	0.903	0.785	57%
Blend: LightGBM + Logistic regression	0.901	0.783	53%
Logistic regression	0.891	0.784	55%
CatBoost	0.895	0.774	51%
LightGBM	0.885	0.772	52%
XGBoost	0.884	0.769	48%
HistGradientBoosting	0.878	0.772	51%
Days since last order	0.866	0.726	53%

- **Gain.** Extra trees beats LightGBM by +0.026 AUC (90% interval +0.011 to +0.040) on 1,314 unseen outlets, 103 of which went quiet. LightGBM itself beats the days-since-last-order rule by +0.046 AUC, but its riskiest 20% catches no more quiet outlets than the rule; Extra trees does (59% vs 53%).
- The site's own churn label and test period differ from ours (its rule scores 0.68, ours 0.73), so its 0.77 is not directly comparable; expect a similar gain of about +0.02 to +0.03 AUC. New scores for every outlet as of 20 Oct are in data/churn_scores_v2.csv.

7. Cluster adoption

Will an outlet start buying from a sales team it has never bought from? We trained on adoptions up to 31 Jul and tested on August - October adoptions at held-out outlets, for the five clusters sold under distributor 1000 (Personal Care sells under its own outlet codes, which the site merges by location and name).

Cluster adoption: our time-honest copy of the test (Aug - Oct adoptions, held-out outlets)



3,885 held-out outlet x cluster pairs, 220 adoptions; Personal Care left out. The site reports 0.90 pooled AUC and 65% in the top 10%.

Model	Validation: AUC within cluster	Test: AUC pooled	Test: AUC within cluster	Test: adoptions in top 10%
Blend: Logistic regression + XGBoost (picked)	0.694	0.791	0.750	32%
XGBoost	0.674	0.809	0.736	35%
Logistic regression	0.693	0.744	0.741	26%
CatBoost	0.671	0.784	0.718	29%
LightGBM	0.671	0.784	0.708	25%
Channel popularity	0.620	0.735	0.649	30%
What neighbours buy	0.481	0.463	0.448	9%

Check before the demo. The site reports a pooled AUC of 0.90 and 65% of adoptions in the top 10%. With a split that only trains on adoptions before the test window, no model we tried gets above 0.81 pooled or 35% in the top 10%, and LightGBM alone catches fewer adoptions in its top 10% than ranking by channel popularity. If the site model trained on August - October adoptions of other outlets, it has seen the test period: re-run it on adoptions before 1 August. Within one cluster its own AUC is 0.67 to 0.86; show that next to the pooled 0.90, which mostly reflects that Unilever HPC adopts far more often than the other teams.

8. Where a better model does not help

Most of the SAR 3.45M rests on benchmarks, simulations and rules. A different algorithm cannot move these; only better data can.

Page	What the figure is	What limits it
Van sales (fleet SAR 1.39M)	Re-zoning simulation and top-quartile drops per van-day	Straight-line distances from outlet GPS; no van tracks or cost per van-day
Warehouse labour, P&L (SAR 639K)	Benchmark against the best quarter of shifts, Monte Carlo range	No pay rates, roster or attendance
Shift supervisor (198 sessions)	Rules plus an IsolationForest anomaly model	No attendance, so the flags cannot be scored at all
Receivables, Collection list (SAR 174K)	Rules on a rebuilt ageing	No payment history to train a model on
Returns (SAR 315K), price (SAR 397K)	Measured rates against cluster medians	Descriptive; promotion and trade-spend data would sharpen them
Cluster harmonisation (SAR 321K)	Own-best neighbourhood benchmark	The lever does not use the adoption model's accuracy

Lever (a year unless stated)	Site base	After	Basis
ML forecast to safety stock	SAR 65.2K	SAR 69.8K	Same bias. SAR 42.6K if forecasts must be unbiased
Inventory cash released (one-off)	SAR 296K	SAR 317K	Same bias. SAR 193K if unbiased
Whitespace / next-best SKU	SAR 104.5K	SAR 105.3K	Scaled by top-5 precision (the list the value uses)
Churn early-warning and win-back	SAR 47.3K	up to SAR 53.4K	Upper bound: whole lever scaled by churners caught
Annual total	SAR 3.45M	SAR 3.466M	All model upgrades together

9. What more data would unlock

Model or page	Data to ask for	What it would change
Demand forecast	12-24 months of daily sales	Learns Ramadan, Eid and the summer peak; allows test windows long enough to rank models
Demand forecast	Promotion and trade-loading calendar	The mid-October burst (one SKU at about 24,000 cases in a week) becomes a known input
Forecast, van loads	Stock on hand; van load / unload by SKU	Turns forecasts into orders and loads; separates lost sales from low demand
Next-best SKU	Rep visit logs and what was offered	Separates "not offered" from "refused"; the gap between 12% and the 42% ceiling is mostly here
Churn	Visit logs or geo-fence flags; closure reasons	Tells closed shops from quiet ones
Cluster adoption	More months of sales	242 test adoptions put about +/-0.03 around the AUC; per-cluster AUCs rest on a handful
Collection list	Customer ledger with payment dates; real credit limits	Allows a payment-risk model; explains the SAR 336K unmatched
Shift supervisor, labour	Attendance, pay rates, roster	Lets the badge flags be scored; moves SAR 639K from Simulated to Corrected

Appendix: files and definitions

File (06_Model_Review/data)	What it holds
before_after_figures.csv, before_after_figures.json	Section 2 as data: page, place, text on the site now, replacement text, type, basis, numeric before and after
forecast_by_level.csv	Per level: site result and bias, our baselines, our copy of the site rule, the picked rule, error at the site bias
forecast_all_rules.csv	Every model and combination rule: validation (25 Aug) and hold-out error, per level
next_best_sku_models.csv	Every recommender on the site's test: validation and test precision, top 5, outlets with a hit
churn_models.csv	Every churn model and blend: two validation folds, test AUC, riskiest-20% capture
adoption_models.csv	Every adoption model and blend: pooled and within-cluster AUC, top-10% capture
next_best_sku_picks_v2.csv	Ten picks per outlet as of 20 Oct 2026 from the picked ranker (SKUs sold somewhere in the last 4 weeks only)
churn_scores_v2.csv	Churn risk per outlet as of 20 Oct 2026: Extra trees (v2) beside the LightGBM version (v1)

- **Weekly error (WAPE)** = sum of absolute weekly misses / sum of actual weekly sales, over every series and the 4 test weeks.
- **Bias** = forecast total / actual total - 1. Negative means forecasting too low.
- **AUC** = how often the model ranks a case that happened (an outlet that went quiet, a cluster adopted) above one that did not; 0.5 is a coin toss.
- **Precision@10** = share of the 10 recommended SKUs that the outlet bought in the 8 weeks after the list was made.
- **90% interval** = paired bootstrap over outlets (2,000 resamples) of the gain against the current model.
- Scripts: 06_Model_Review/scripts (fc_race, fc_eval, reco_race, reco_final_v2, reco_picks_v2, churn_race, churn_final_v2, adoption_race, charts, data_changes, build_report, check_report), run in that order with the 05_Predictions work files.