

Margin & Exception Intelligence: Full Analysis

Halbrook Industrial Supply

What changed that should not have changed—and what deserves management attention?

Invoice lines analysed	Blended margin decline	Issues for management	Seeded problems found
278,169	1.59 pts	20	7 of 7

Demonstration project using synthetic data designed to simulate a realistic operating company.

1. The business problem

Halbrook Industrial Supply grew revenue to \$38.40M and lost 1.59 points of gross margin over two years. The monthly reporting package showed both facts and explained neither. It reported one blended margin, in absolute dollars, so a decline spread across twenty-four months never looked like an event and nothing in it said which accounts, which products or which branches had moved.

The company has 142 accounts and 278,169 invoice lines across 36 months. Somewhere in it are the reasons. The question is not whether the data holds them — it is whether anyone can be pointed at the few that matter without reading all of it.

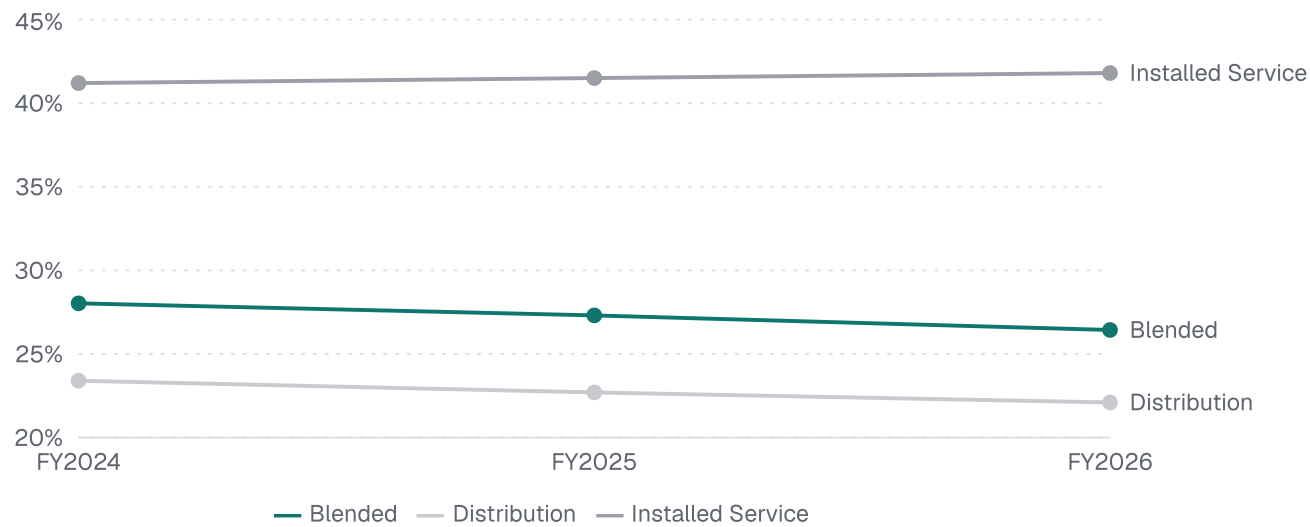
2. Why standard reporting is not enough

A blended margin is an average of two businesses. Distribution sells parts at 22.1%; Installed Service sends technicians out at 41.8% — nearly double. Both moved a little: Distribution gave up 1.30 points, Installed Service gained 0.60. Weighted, that is 0.81 points of the fall. The other 0.79 is Installed Service dropping from 26% of revenue to 22% — the same margins, on a worse blend.

That is a different problem from either line getting worse, and it needs a different decision. Pricing changes alone would not reverse the mix component; holding segment margins constant, a higher share of Installed Service revenue would mechanically improve the blend, on a longer timescale. A report that shows only the blend cannot tell the two apart.

One line earns nearly double the other’s margin

Gross margin by line of business and for the company. **Teal** is the blended margin — the figure the monthly reporting package carried. The two lines are far apart and both moved a little: Distribution gave up 1.30 points and Installed Service gained 0.60. Weighted by what each was worth, that accounts for about half of the blended fall of 1.59 points. The rest is the mix.



The decline, decomposed

Each segment's own margin change and its change in share, summing to the movement exactly. There is no residual, so there is nothing for anyone to explain away.

Driver	Effect	What it means
Net effect of within-line margin changes	-0.81 pts	one line earning less on what it sold, the other a little more
Mix shifting toward Distribution	-0.79 pts	the same margins, on a worse blend of revenue
Total decline	-1.59 pts	what the blended margin actually did

3. Analytical approach

Rules run before statistics, because a rule management already understands produces an explanation rather than a flag. A line billed below what it cost is an economic exception on its face, even where the

business later decides it was intentional. An account two points past its segment’s discount schedule is a conversation someone can have. Neither needs a reader to know what a robust deviation is.

Statistical detection then covers what no rule anticipated — a segment sitting away from its peers, or a series that changed level on a date. Attribution runs alongside both, because a movement is only actionable once it is attributed, and at account level attribution is the route that works: it never asks whether an account is unusual, only how much of the movement it accounts for, which has an exact answer.

Every measure, and what it is held against

A level carries the season, so it is compared with the same period a year earlier; a rate does not, so it is compared with the period before. The definitions ship with the figures rather than being restated beside them.

Measure	Compared with	Across segments
revenue	same period last year	Own history only
gross margin	prior period	Comparable
quantity	same period last year	Own history only
discount rate	prior period	Comparable
unit cost	prior period	Own history only

4. What the solution found

629 detections across six ways of cutting the book, aggregated to 122 rows, of which 20 are promoted to management. Each carries what it departed from, by how much, what it is estimated to cost, and who should act — and the action follows from the evidence rather than from the measure, so a concession past a configured schedule is renegotiated while the same measure away from its peers is a question.

The management tier, most severe first

Severity weighs the estimated size, the detector's measured precision, and how many periods the condition held. Confidence is measured against the register rather than asserted — a detector nobody scored is carried as unmeasured. The estimates are not additive: an account sits inside a branch and a family inside a line of business, so the same economics appears at every cut. They size findings; they do not sum to a total.

Where	Grain	Measure	Est. impact	Action
179 lines	Invoice line	gross profit	-\$14K/yr	Correct the record
Omaha	Branch	gross margin	-\$210K/yr	Investigate
Electrical	Product family	cost price spread	-\$292K/yr	Reprice
Pumps	Product family	cost price spread	-\$101K/yr	Reprice
Trellis Works	Account	discount rate	-\$45K/yr	Renegotiate terms
Kestrel Fabrication	Account	discount rate	-\$37K/yr	Renegotiate terms
Larkspur Processing	Account	discount rate	-\$20K/yr	Renegotiate terms
Trellis Mechanical	Account	cost price spread	-\$20K/yr	Reprice

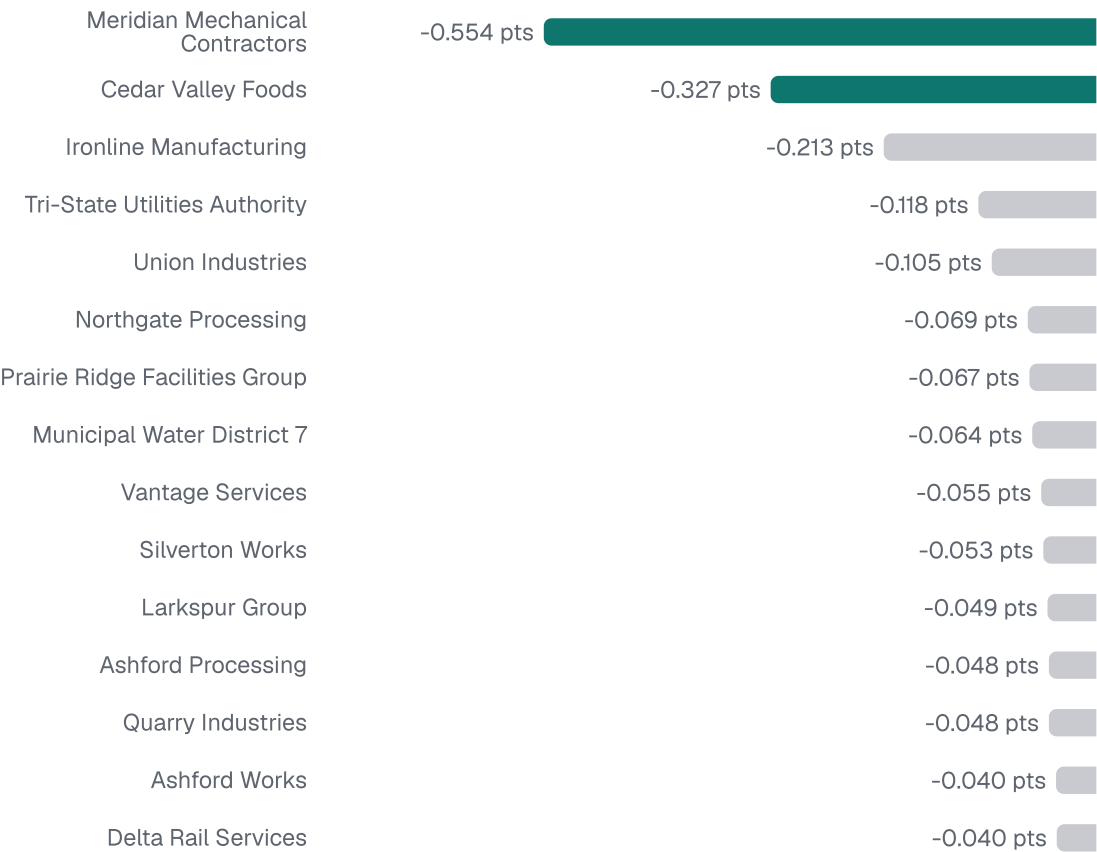
One issue, in full

Every row reads like this. The sentence is built from the row's own fields, so it cannot disagree with the numbers beside it.

179 records: gross profit of -\$234.20 against \$0.00, \$234.20 from the policy threshold. Estimated cost: \$41,923 of gross profit. Action: correct - the record is wrong, so it is corrected rather than negotiated.

2 accounts explain half of the decline

Each account’s contribution to the movement in blended margin, in points. The contributions sum to the movement exactly, so this is an attribution rather than a ranking of how unusual anything looks. **Teal** marks the accounts that together reach half of it. Every account here contributed to the decline; none offset it.



Concentration: the thing that was not there

Reyes expected dependence on a few accounts to be part of the story. It is not: concentration falls on every measure, including one computed across the whole distribution rather than the top ten. A demonstration that only ever confirms what it looks for has not shown it can tell the difference.

Measure	Opening	Closing	Direction
top 1 share	9.3%	7.8%	falling
top 5 share	28.2%	25.3%	falling
top 10 share	42.2%	40.3%	falling
herfindahl	2.5%	2.2%	falling

5. Management use

What went wrong is not the same as what someone must do, so every issue carries an action and the actions go to different people. A cost that rose and was not passed on needs repricing. A concession past the configured threshold is routed for renegotiation review. Volume leaving an account at an unchanged margin is a demand question, not a pricing one. A mis-billed line needs correcting, and nobody needs to negotiate anything.

On this dataset the most severe row for 6 of 6 seeded problems carries the action the register says was needed. That is the strict reading, and it is the one worth publishing. A weaker question — whether the right action appeared anywhere among the rows for a problem — answers 6 of 6, because five of the six produce more than one action and being right somewhere is not the same as leading with it. The mix shift is the one to read carefully: it is routed to monitor rather than investigate, because a shift between lines is not recovered by chasing anybody.

Where each seeded problem went

Scored against what the register says should have been done about it. Led with is the action on the problem's most severe row, which is what a reader sees first. Routed as required is the share of the problem's rows carrying the needed action. One figure here used to mean only that the right action appeared somewhere among them, which is a weaker thing and read as a stronger one.

Problem	Needed	Led with	Routed as required	Also routed to
account cost not passed through	Reprice	Reprice	60%	1
family cost not repriced	Reprice	Reprice	50%	1
discount above threshold	Renegotiate terms	Renegotiate terms	47%	1
silent volume attrition	Investigate	Investigate	100%	nothing else
mix shift	Monitor	Monitor	50%	1
branch margin divergence	Investigate	Investigate	100%	nothing else

6. Methods and technical note

Every accuracy figure in this document is measured against a ground-truth register, not asserted. The dataset is generated with its problems deliberately seeded and recorded, so what the engine found can be checked against what was put there. The register is withheld from the engine and used only after the outputs exist, which is the independence that matters — it does ship beside those outputs in the evidence package, because an accuracy figure whose evidence is absent cannot be checked at all.

The evidence package ships a verifier that rebuilds every published figure from the data in the package, using nothing but the standard library. **95 of 95 checks pass on this build** — across both demonstrations, and nothing in the exception headline is declared rather than derived. Run it yourself: `python verify.py`.

Every seeded problem, and whether it was found

A problem is credited only to a detection at the grain the register records it at. A break is dated too, because a detector that reports only that something changed cannot be held to when.

Problem	Grain	Found	Break dated
account cost not passed through	Account	6 of 6	-1 mo
family cost not repriced	Product family	2 of 2	0 mo
discount above threshold	Account manager	1 of 1	—
silent volume attrition	Account	3 of 3	+15 mo
mix shift	Line of business	1 of 1	+12 mo
concentration growth	Account	Correctly absent	—
branch margin divergence	Branch	1 of 1	—

180 of 180 mis-billed lines were found, and 6 of 6 segment-level problems. The one problem deliberately not in the data was not reported. 14 of 14 individual segments were identified. The ones missed are the smallest members of the affected sets, whose contribution to the decline is correspondingly small.

Coverage is only half of the question, and it is the flattering half: a detector can find every planted problem by flagging a great deal. The other half is how much irrelevant work the queue creates. Two figures answer it, because they are not the same question. A row lands on a seeded population if its segment carries a registered condition at that row's grain — 32% of the whole queue does. Issue precision is the question a manager has when they open a row, and it is measured against what the register says each condition *is*: the measure whose movement defines it, a detector that can establish it, the grain the evidence appears at, and the action it asks for. All four have to agree, and that is 19%. The gap between them is rows that found the

right account and described a different thing about it — 9% of the queue is real movement on a seeded population that is not the registered finding, and it is reported as corroboration rather than as either a finding or noise.

This matters more than it sounds. Installed Service carries the mix shift, and a trend break fires on that line's revenue, gross profit, realized unit price, order value and revenue share. Scoring on the segment alone credited all five, published a trend-break precision of 71% at line-of-business grain, and used it to promote a realized-price movement above the mix shift itself. Precision is measured per detector, grain and measure now — the same lesson as the grain, one level down.

So the queue is tiered, and only one tier is put forward as work. 20 rows reach the management tier — on decisive evidence, or because the route that found them has measured precision of at least 70% at that grain and on that measure — and they run at 100% issue precision *in sample*. The other 102 are review signals: explained, checkable, and not presented as something to act on. Calling all 122 of them issues was a promise these same figures disproved.

In sample, and the qualification is not a formality. A route is promoted on precision measured against the same register the tier is then scored against, so the threshold is calibrated on this demonstration's own answer key. Held out one condition at a time, 5 of the 6 promoting routes make no detections anywhere except on the condition they promote — a precision of 1.00 for the discount rule means the eight accounts it flagged are the eight the register seeds, with nothing else in the data for it to be wrong about. Splitting this population into calibration and evaluation sets would not fix that: one company carries one instance of each condition, so holding the instance out leaves nothing to evaluate. What the tier shows is the workflow — measure a route's quality, promote the reliable ones, leave the weak ones as review signals — not a generalisation estimate.

That precision is bought with recall, and the price is published beside it. The whole queue carries a row for 7 of the 7 seeded conditions; the promoted tier carries 6. The one it leaves behind is the silent volume attrition, whose route measures 27% precision — the three real accounts rank first, second and third, and eight other accounts shrank too, so no threshold separates them. Those rows sit in the review tier with their evidence attached rather than being promoted on a measurement that does not support it.

How much of the queue is real

Two measures. A row is on a seeded population if its segment carries a registered condition at that row's own grain; it is a correct issue only if it also carries the action that condition needs. Deliberately strict about grain: a branch row gets no credit because a seeded customer happens to sit in that branch. On a client engagement there is no register and so no measurement, which is a property of client work rather than a gap here.

Rows reviewed	Rows	On a seeded population	Stated problem is the problem
top 10	10	100%	100%
top 20	20	100%	100%
top 60	60	45%	38%
all	122	32%	19%

How often each route fires on something clean

A cohort of accounts carries no seeded condition and absorbs none of the compensation. They are unseeded account-level controls rather than accounts nothing touched: seven sit in the branch the register seeds a divergence on, so a flag on one is a false positive at account grain and not proof that nothing around it moved. Lift is that rate against how often those accounts simply occur in the book: a route at 1.0 is indifferent to whether an account is clean. A null means the cohort is not held out at that grain, so nothing was measured — which is not the same as no false positives.

Route	Grain	Flags	Lift on clean
account cost outran price	Account	6	0.00
cost outran price	Product family	2	Not measured
discount above authority	Account	8	0.00
line below cost	Invoice line	180	0.00
line priced below group	Invoice line	179	0.00
peers departure	Account manager	11	Not measured
peers departure	Branch	3	Not measured
peers departure	Account	28	0.18
peers departure	Customer segment	6	Not measured
trend break	Account manager	19	Not measured
trend break	Branch	13	Not measured
trend break	Account	118	0.73
trend break	Line of business	7	Not measured
trend break	Product family	28	Not measured
trend break	Customer segment	10	Not measured
volume attrition	Account	11	0.46

4 of the 7 measurable routes never fired on a held-out account. The rest are reported as they measure, because a detector that flags clean accounts at the rate they occur in the book is not avoiding them, and saying so is more useful than a figure that implies otherwise.

How wrong the size estimate is

The estimate uses only what is in the data: the gap a detector measured, times the exposure that gap applies to. The register's magnitude is a counterfactual against an unseeded ledger, which nothing outside a demonstration can have. Comparing them is the only way a claim about financial impact is measured rather than asserted. These are the conditions carrying a registered financial magnitude, and the median error is over them - not over every sized row in the queue, which has no ground-truth amount to be scored against at all.

Problem	Actual, per year	Estimated	Ratio
account cost not passed through	\$25K	\$83K	3.33x
family cost not repriced	\$188K	\$581K	3.09x
discount above threshold	\$66K	\$79K	1.20x
silent volume attrition	\$113K	\$235K	2.08x
branch margin divergence	\$123K	\$210K	1.71x

The median estimate came in at 2.08x the seeded magnitude and the worst at 3.33x, on account cost not passed through. It runs high because a comparison against peers attributes all of a segment's difference to the problem, when some of it is the segment being genuinely different. That is what the figure is worth, and publishing it beats calling the estimate accurate.

Generated 2026-09-14 from 278,169 invoice lines. Every figure in this document is read from the published artifacts; none is written by hand.