THE METASCIENCE **Volume 4, Issue 4 (Oct-Dec)-2026**
(An International Multidisciplinary Research Journal)

Beyond Point Accuracy: Evaluating Probabilistic Machine-Learning Cash-Flow Forecasts for Small and Medium-Sized Enterprises

Nikolić Mariam [1], Stefan Anna [2]
[1] University of Zambia, ZAMBIA
[2] University of Botswana, BOTSWANA

Keywords

Cash-Flow Forecasting, SMEs, Probabilistic Forecasting, Machine Learning, Conformal Prediction, Forecast Evaluation, Financial Planning and Analysis, Decision Support

ABSTRACT

Machine-learning cash-flow forecasting is increasingly offered to small and medium-sized enterprises (SMEs) as part of AI-driven financial planning and analysis [1]. Most evaluations of such tools report only point accuracy, yet SME owners use forecasts to answer a different question: how likely is the business to run short of cash, and how early can it know? This article proposes an evaluation framework that judges SME cash-flow forecasts on three dimensions together: scaled accuracy, calibration of prediction intervals, and decision value measured through cash-shortfall alerts. We demonstrate the framework in a reproducible experiment on 300 simulated SMEs with weekly cash flows over three years, comparing a naive baseline, a seasonal-naive baseline, a gradient-boosted quantile model and the same model with conformal calibration. The machine-learning model matched but did not beat the seasonal-naive baseline on point accuracy (relative mean absolute error 0.39 for both, against 1.00 for the naive method). Its raw 80 percent prediction intervals covered only 61 percent of outcomes; conformal calibration raised coverage to 73 percent, with the remaining gap traced to demand shocks absent from the calibration window. For shortfall alerts, the naive method detected the most shortfalls but raised false alarms in 25 percent of safe cases, against 6 to 7 percent for the better methods. We then replicate the evaluation on real weekly revenue of 327 small online sellers from a public Brazilian marketplace dataset. There the ranking of methods changed: the seasonal-naive benchmark failed (relative MAE 1.94) because the sellers were still growing, the machine-learning model was most accurate (0.97) and had the lowest pinball loss, its raw 80 percent intervals again covered only 65 percent of outcomes, and conformal calibration restored coverage to 82 percent. A shortfall alert based on the lower 10 percent quantile, which worked in simulation, raised false alarms in most safe cases on these intermittent revenue series. The results show that accuracy alone can mislead SME adoption decisions and that calibration and decision-oriented metrics should be standard in evaluating AI-driven financial planning tools.

Introduction

Cash is the binding constraint for most small businesses. Evidence from transaction data on hundreds of thousands of US small firms shows that the typical business holds only a few weeks of cash buffer and experiences highly volatile inflows and outflows [2]. A forecast that warns an owner several weeks before a shortfall can therefore be worth more than any amount of retrospective reporting. AI-driven financial planning and analysis (FP&A) systems for SMEs promise exactly this: machine-learning cash-flow forecasts combined with risk-aware decision support for technology-enabled operations [1].

How should such forecasts be judged? Forecasting research has long warned that the method that wins on one accuracy measure need not win on another [3][4], and large forecasting competitions have shown that

machine-learning methods do not automatically outperform simple statistical benchmarks [5][6]. For SME cash management, the problem is sharper still. An owner deciding whether to draw on a credit line needs to know the range of plausible outcomes, not just the most likely one. A forecast whose intervals are too narrow will produce false confidence; one whose intervals are too wide will produce alarms that are ignored.

This article builds on the AI-driven FP&A framework for SMEs described in [1] and addresses the evaluation question it raises. We make three contributions. First, we propose a three-dimension evaluation framework covering accuracy, calibration and decision value. Second, we demonstrate the framework in a reproducible simulation of 300 SMEs, comparing two classical baselines with a gradient-boosted quantile model, with and without conformal calibration. We then replicate the evaluation on real revenue data from 327 small online sellers, which tests whether the simulation's conclusions survive contact with real SME cash inflows. Third, we draw practical guidance for SME owners, software providers and lenders on how forecast tools should be evaluated before adoption.

Background and Related Work

Forecast Accuracy Measures

Percentage errors are unreliable when the forecast quantity can be close to zero or negative, which is common for net cash flow. Scaled measures avoid this by comparing a method's error with that of a simple benchmark on the same series [3]. We follow this logic and report mean absolute error relative to a naive method, so that a value below one indicates improvement over the benchmark. Rolling-origin evaluation, in which forecasts are produced from many successive origins and errors are pooled, gives a more reliable picture than a single hold-out split [4][7].

Probabilistic Forecasts and Calibration

A probabilistic forecast should be both calibrated, meaning that stated probabilities match observed frequencies, and sharp, meaning that intervals are as narrow as calibration allows [8][9]. Proper scoring rules such as the pinball (quantile) loss reward both properties together [10]. Quantile regression provides a direct way to estimate conditional quantiles [11], and gradient boosting extends this to flexible non-linear models [12]. Quantile models trained in this way are not guaranteed to be calibrated out of sample. Conformal prediction offers a distribution-free correction that guarantees coverage when calibration and test data are exchangeable [13], and conformalized quantile regression applies this idea to adjust the intervals of a quantile model [14].

Machine Learning in SME Financial Planning

The AI-driven FP&A approach in [1] combines machine-learning cash-flow forecasting with risk-aware decision support. Its premise is that SMEs can now access forecasting capability once limited to firms with dedicated finance teams. That premise makes rigorous evaluation more important, not less: an SME owner rarely has the expertise to question a model's output, so the evaluation standards applied by software providers and advisers effectively determine whether forecasts help or mislead. Table 1 summarizes the evaluation dimensions we propose.

Table 1: Evaluation Dimensions for SME Cash-Flow Forecasts

Dimension	Question for the SME owner	Metric used here	Ideal value
Accuracy	How close is the central forecast to what happens?	Mean absolute error relative to naive method [3]	Below 1.00

Dimension	Question for the SME owner	Metric used here	Ideal value
Calibration	Can I trust the stated range of outcomes?	Coverage of the 80% prediction interval [8]	Close to 80%
Overall probabilistic skill	Is the whole forecast distribution useful?	Mean pinball loss across quantiles [10]	As low as possible
Decision value	Will I be warned before I run short of cash?	Shortfall detection and false-alarm rates	High detection, low false alarms

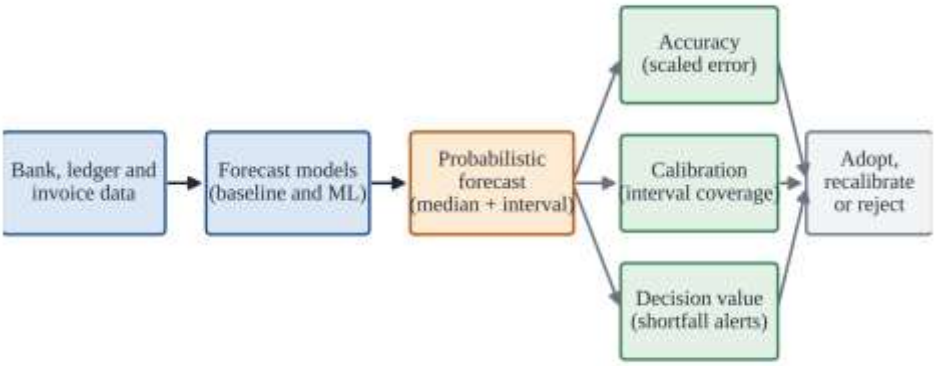
Human Judgment and Trust in Forecasts

Forecasts in small firms are rarely used exactly as produced. Owners and accountants adjust them using knowledge of contracts, customers and seasonal events, and such judgmental adjustments can either add value or introduce bias depending on their size and direction [15]. Trust is also fragile: people tend to abandon an algorithm after seeing it err, even when it outperforms human judgment [16], although they are more willing to keep using an imperfect algorithm when allowed to modify it slightly [17]. For SME cash forecasting, this implies that honest communication of uncertainty is not a technical refinement but a condition for sustained use. A forecast whose intervals are well calibrated sets realistic expectations, so that errors within the stated range do not destroy trust.

Evaluation Framework

Figure 1 shows the framework. Bank, ledger and invoice data feed one or more forecasting models. Each model must produce a probabilistic forecast, at minimum a median and an interval, rather than a single number. That forecast is then judged on the three dimensions in Table 1, and the combined evidence supports a decision to adopt the model, recalibrate it or reject it in favour of a simpler benchmark.

Figure 1 *Three-dimension evaluation framework for SME cash-flow forecasts*



Source: authors' design.

The decision-value dimension links forecasts to the action an owner actually takes. We define a cash-shortfall event as a cash balance that would fall below zero at the end of the forecast horizon. A method raises an alert when its lower 10 percent quantile, added to the current balance, falls below zero. Detection rate is the share of true shortfalls that are preceded by an alert; false-alarm rate is the share of safe cases in which an alert is raised. These two rates make the cost of a poorly calibrated forecast visible in business terms.

Experimental Design

To demonstrate the framework reproducibly, we simulated weekly net cash flows for 300 SMEs over 156 weeks. Each firm was given its own scale, trend, annual seasonality with random phase and amplitude, a monthly payroll and rent outflow, idiosyncratic noise and, for about one-third of firms, a temporary demand shock of four to eleven weeks during the third year. Initial cash balances were set at three to twelve typical weeks of gross flow. The data are synthetic and are intended to resemble the volatility patterns reported for small businesses [2]; they do not represent any real firm.

The forecasting target was cumulative net cash flow over the next 13 weeks, a common short-term treasury horizon, scaled by each firm's typical weekly flow. Table 2 lists the four methods compared. The gradient-boosted quantile model was trained as a single global model across all firms using lagged flows, rolling means, the same 13 weeks of the previous year, seasonal terms and the current balance as features, implemented in scikit-learn [18].

Table 2: Forecasting Methods Compared in the Experiment

Method	Point forecast	Prediction interval
Naive	Mean of the last four weeks multiplied by the horizon	Empirical residual quantiles from training period
Seasonal naive	Sum of the same 13 weeks in the previous year [19]	Empirical residual quantiles from training period
ML quantile model	Gradient-boosted median [12][11]	Gradient-boosted 10th and 90th percentiles
ML + conformal	Same as ML quantile model	Quantile interval widened by conformal correction [14]

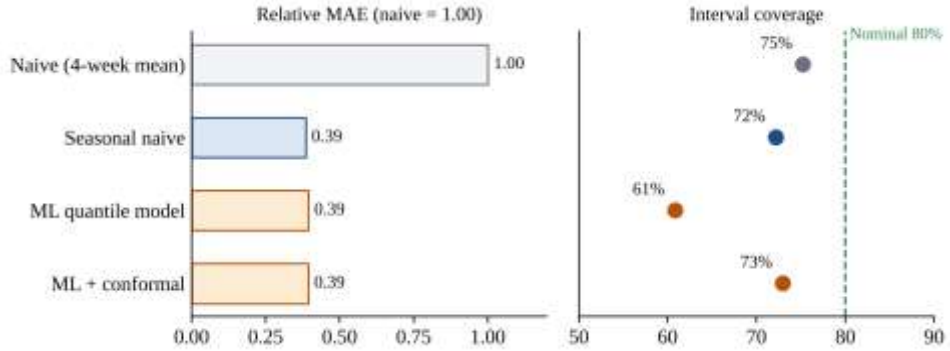
Data were split by time. Forecast origins in weeks 60 to 79 formed the training set (6,000 cases), origins in weeks 80 to 90 the conformal calibration set, and origins in weeks 104 to 142 the test set (11,700 cases). Of the test cases, 1,705, or 14.6 percent, were shortfall events. The conformal correction was the empirical quantile of interval-violation scores on the calibration set, chosen to target 80 percent coverage [14]. A fixed random seed makes every figure reproducible.

Results

Accuracy and Calibration

Figure 2 shows the accuracy and calibration results. Both the seasonal-naive method and the machine-learning model reduced mean absolute error by about 61 percent relative to the naive method, to a relative value of 0.39. The machine-learning model did not improve on the seasonal-naive baseline. This mirrors findings from large forecasting competitions, in which simple seasonal benchmarks remain hard to beat when the dominant pattern is regular seasonality [5].

Figure 2 *Relative accuracy and 80% interval coverage by method (synthetic data)*



Source: authors' simulation.

The calibration results separate the methods more clearly. The raw quantile model's 80 percent intervals covered only 60.9 percent of outcomes, meaning that owners would be surprised far more often than the forecast implied. Conformal calibration raised coverage to 73.0 percent. Coverage did not reach the nominal 80 percent because the test year differed from the calibration window, most notably through demand shocks that hit about one-third of firms at different points in the year, so the exchangeability assumption behind conformal guarantees did not hold [13]. This is itself a practical lesson: calibration must be monitored over time, not established once.

Table 3: Summary of Forecast Evaluation Results on the Test Set (Synthetic Data)

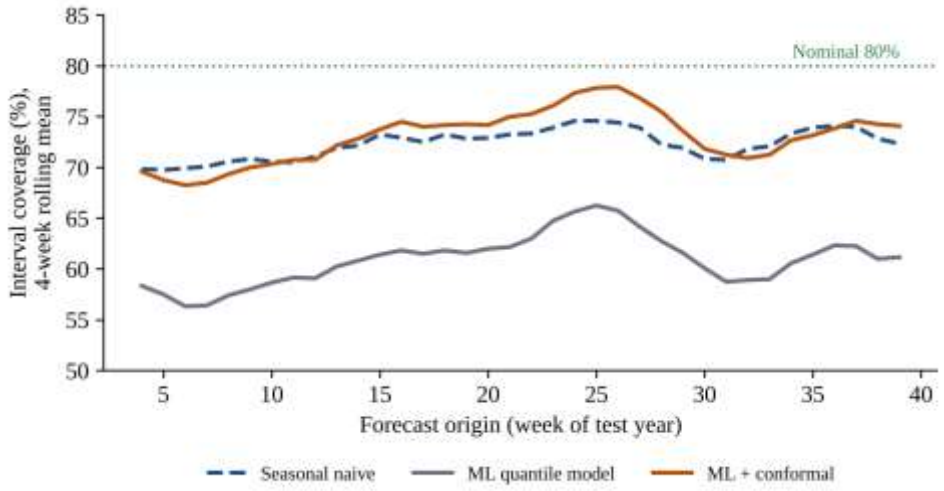
Method	Relative MAE	80% interval coverage	Mean pinball loss	False-alarm rate
Naive	1.00	75.2%	3.60	25.1%
Seasonal naive	0.39	72.2%	1.51	6.8%
ML quantile model	0.39	60.9%	1.61	4.6%
ML + conformal	0.39	73.0%	1.55	6.2%

Note: relative MAE is mean absolute error divided by that of the naive method. Pinball loss is averaged over the 10th, 50th and 90th percentiles in units of typical weekly flow.

Calibration Over Time

Average coverage can hide periods in which a forecast is much less reliable. Figure 3 tracks interval coverage across the test year using a four-week rolling mean. The raw quantile model stayed between about 56 and 66 percent throughout, well below the nominal level. The conformal and seasonal-naive intervals moved together between about 68 and 78 percent, approaching the target in the middle of the year and falling back as demand shocks affected more firms. A provider that monitored only an annual average would have missed this variation.

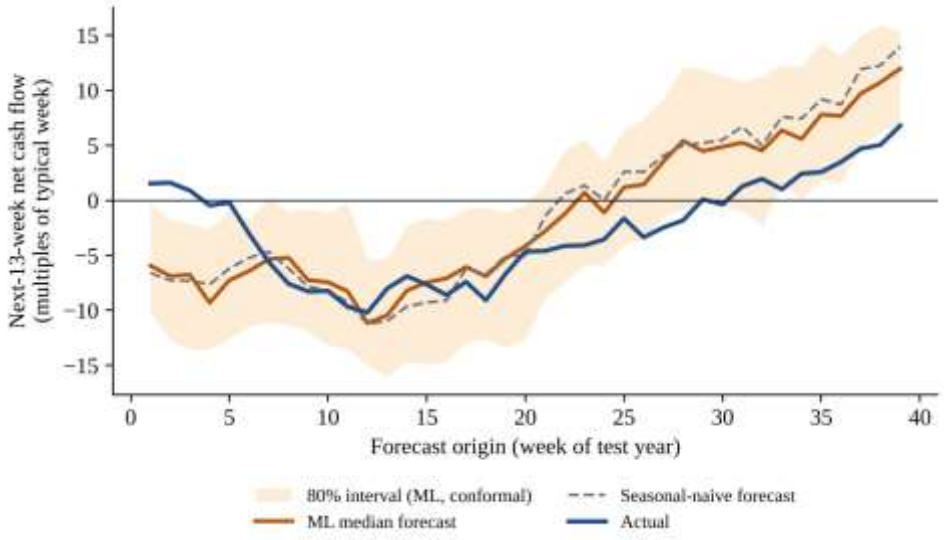
Figure 3 *Rolling 80% interval coverage across the test year (synthetic data)*



Source: authors' simulation.

Table 3 summarizes the results. The pinball loss confirms that the seasonal-naive method had the best overall probabilistic skill in this experiment, with conformal calibration bringing the machine-learning model close behind. Figure 4 illustrates the forecasts for one simulated firm during the test year, showing how the conformal interval widened during a period of rapid change.

Figure 4 *Example 13-week cash-flow forecasts for one simulated firm across the test year*



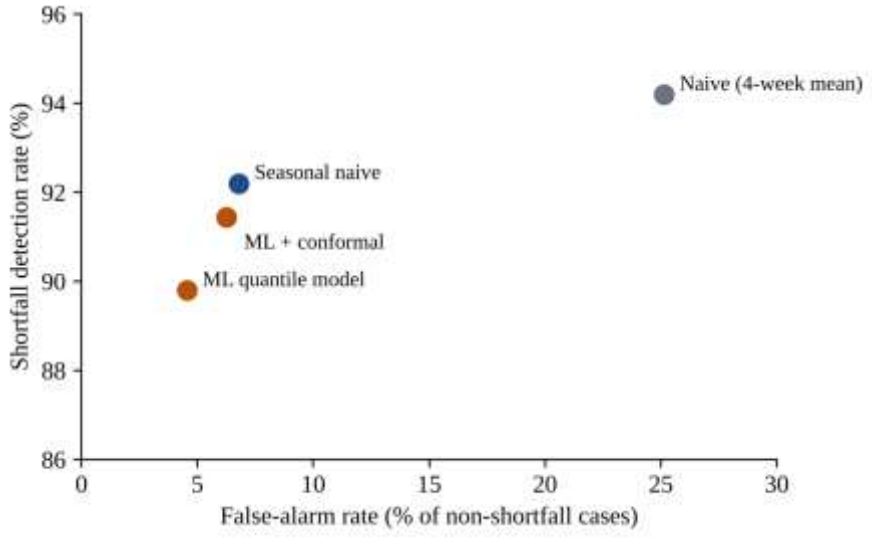
Source: authors' simulation.

Decision Value

Figure 5 places each method on the trade-off between detecting shortfalls and avoiding false alarms. The naive method detected the most shortfalls, 94.2 percent, but only because its wide and poorly centred forecasts raised alerts in a quarter of all safe cases. An owner receiving false alarms that often would soon learn to ignore them. The seasonal-naive method detected 92.2 percent of shortfalls with a false-alarm rate of 6.8 percent. The raw machine-learning model produced the fewest false alarms, 4.6 percent, but its

overconfident intervals missed more shortfalls, detecting 89.8 percent. Conformal calibration moved it to 91.4 percent detection at a 6.2 percent false-alarm rate.

Figure 5 *Shortfall detection versus false-alarm rate by method (synthetic data)*



Source: authors' simulation.

Replication on Real SME Revenue Data

Data and Design

Simulation controls the data-generating process but cannot show how a method behaves on real small-business flows. We therefore repeated the evaluation on the public Olist dataset of orders placed on a Brazilian online marketplace between 2016 and 2018[20]. Each order item records the seller, the price and freight, and the purchase time. Weekly revenue was computed for every seller from 2 January 2017 to 19 August 2018 (85 weeks). We retained the 327 sellers that were trading from the start of the panel and had sales in at least 15 of the first 52 weeks. These are genuinely small businesses: median weekly revenue was BRL 134 (interquartile range BRL 74 to 295), and many weeks had no sales at all.

Three adaptations were needed. The data record revenue but not outflows or bank balances, so the target was cumulative revenue over the next four weeks, scaled by each seller's typical weekly revenue. The panel is too short for a 13-week horizon with a full year of seasonal history. And because there is no balance, a shortfall was defined as four-week revenue below half of the seller's trailing 12-week run-rate, a sharp fall in cash inflow; 36.1 percent of test cases met this definition. The four methods of Table 2 were otherwise unchanged. The quantile model was a single global LightGBM model[21] with lagged revenue, rolling means and volatility, the share of zero weeks, the same weeks of the previous year, week-of-year terms and seller age as features. Forecast origins in weeks 20 to 51 formed the training set (10,464 cases), weeks 52 to 61 the conformal calibration set (3,270), and weeks 62 to 80 the test set (6,213). Alerts used the rule of the Evaluation Framework section (lower 10 percent quantile below the shortfall threshold) and, as a comparison, a rule based on the median forecast.

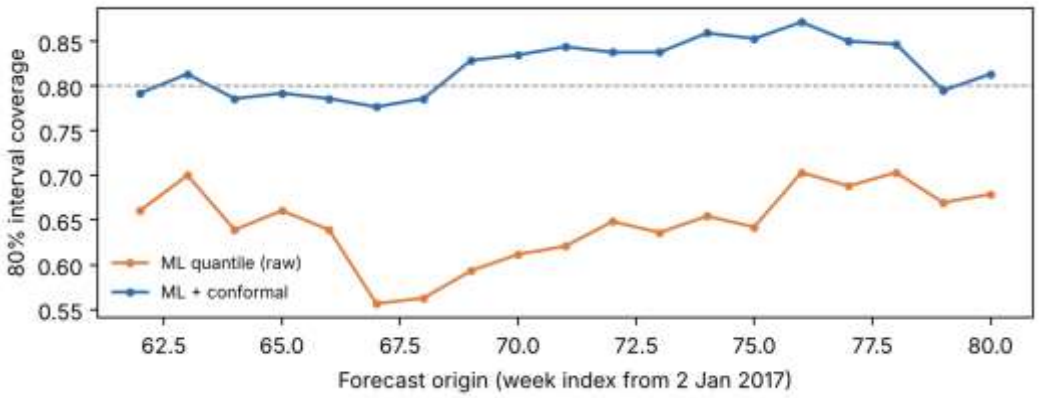
Results

Table 4: Forecast Evaluation on Real Seller Revenue (Olist, 327 Sellers, 6,213 Test Cases)

Method	Relative MAE	80% coverage	Mean interval width	Pinball loss	Alerts, lower-quantile rule (detection / false alarm)	Alerts, median rule (detection / false alarm)
Naive	1.00	86.8%	10.83	0.915	91% / 90%	46% / 15%
Seasonal naive	1.94	73.9%	15.41	1.581	80% / 76%	28% / 25%
ML quantile model	0.97	64.6%	4.64	0.827	93% / 69%	30% / 25%
ML + conformal	0.97	82.1%	5.93	0.814	99% / 98%	30% / 25%

Table 4 reports the results. Two findings from the simulation held and one reversed. The reversal concerns point accuracy: the seasonal-naive method, best in simulation, was the worst here (relative MAE 1.94). Most sellers grew rapidly during 2017 as the marketplace expanded, so the same weeks a year earlier were a poor guide. The machine-learning model was the most accurate (0.969), though only slightly better than the naive method, and it had the lowest pinball loss. A buyer comparing tools on a single benchmark, or on results from a different kind of business, would draw the wrong conclusion in one setting or the other, which is the argument for evaluating each candidate tool on the user's own data.

Figure 6 Rolling 80% Interval Coverage of the Machine-Learning Model on Real Seller Revenue, Raw and Conformal



Source: authors' analysis of the Olist public dataset.

The calibration finding held. Raw quantile intervals covered only 64.6 percent of outcomes against a nominal 80 percent and stayed between 56 and 70 percent at every test origin (Figure 6). Conformal calibration raised coverage to 82.1 percent, between 78 and 87 percent across origins, while keeping intervals far narrower than those of the benchmarks (mean width 5.93 against 10.83 for the naive method). The decision-value finding held in a sharper form. Because revenue for many small sellers is intermittent, the lower 10 percent quantile was close to zero for most of them, so the lower-quantile alert fired in most cases whether or not a shortfall followed: 98 percent false alarms for the conformal model and 90 percent for the naive method. A median-based rule was usable but blunt, detecting 46 percent of shortfalls with 15 percent false alarms for the naive method and 30 percent with 25 percent for the machine-learning model. No method gave an owner a reliable

early signal of a sharp revenue fall four weeks ahead. Decision-value metrics exposed this, whereas accuracy and coverage, both acceptable for the conformal model, did not.

Discussion

Three lessons emerge. First, point accuracy alone would have suggested that the machine-learning model and the seasonal-naive method were interchangeable, yet their intervals behaved very differently. An SME tool marketed on accuracy statistics alone could therefore present an overconfident forecast as equally reliable. Second, the overconfidence of the raw quantile model translated directly into missed shortfalls, the outcome an owner most needs to avoid. Third, conformal calibration is a simple and inexpensive remedy, but it depends on the calibration data resembling the future; when conditions shift, coverage degrades and must be monitored. The real-data replication adds a fourth: the alert rule must be designed for the shape of the series, because a rule that works for smooth simulated cash flow can be useless for intermittent revenue, and a benchmark that is strong for mature seasonal firms can be weak for growing ones.

These findings support the risk-aware emphasis of the AI-driven FP&A approach in [1]. A forecasting engine that serves decision support must report and monitor calibration alongside accuracy. Table 5 translates the results into recommendations for the main stakeholders.

Table 5: Practical Recommendations for Stakeholders in SME Cash-Flow Forecasting

Stakeholder	Recommendation
Software providers	Report interval coverage and pinball loss alongside accuracy; always benchmark against seasonal-naive forecasts; apply and monitor conformal or similar calibration.
SME owners and advisers	Ask how often stated ranges have actually contained the outcome; judge alerts by both detection and false-alarm rates.
Lenders using SME forecasts	Treat uncalibrated intervals with caution in credit monitoring; prefer tools that disclose rolling calibration results.
Researchers	Evaluate forecasts on decision-oriented metrics and under distribution shift, not only on stable hold-out periods.

From Forecasts to Cash Buffers

A probabilistic forecast becomes useful when it is translated into an action. The simplest translation is a cash-buffer rule. If B is the current balance and Q10 is the 10th percentile of forecast net cash flow over the horizon, the additional liquidity needed to keep the balance non-negative in nine cases out of ten is:

$$Required\ buffer = max(0, -(B + Q10))$$

An owner can meet this requirement by arranging a credit line, delaying discretionary spending or accelerating receivables. The rule only works if Q10 is calibrated; an overconfident model understates the buffer and leaves the firm exposed exactly when it most needs protection. Table 6 shows how each element of a probabilistic forecast can be read and acted on.

Table 6: Reading a Probabilistic Cash-Flow Forecast in Practice

Forecast element	Interpretation for the owner	Typical action
Median (Q50)	Most likely 13-week net cash flow	Base plan for payments and purchases
Lower bound (Q10)	A bad but plausible outcome, worse than nine cases in ten	Size the cash buffer or credit line
Upper bound (Q90)	A good but plausible outcome	Plan optional investment only if reached
Interval width	How uncertain the next quarter is	Wider interval: review assumptions and contingencies
Shortfall alert	Q10 implies a negative balance	Contact lender early; adjust spending

Reporting Standards for SME Forecast Evaluations

The experiment suggests a minimum reporting standard that would let SME owners, advisers and lenders compare forecasting tools on a fair basis. Rolling-origin evaluation with a clear separation of training, calibration and test periods avoids optimistic results [4][7]. Benchmarks must include seasonal-naive forecasts, which are simple yet strong [19]. Probabilistic results should be reported with proper scoring rules and coverage [10][8]. Table 7 sets out a checklist.

Table 7: Minimum Reporting Checklist for SME Cash-Flow Forecast Evaluations

Item	What to report
Data and horizon	Number of firms, history length, frequency, forecast horizon and target definition
Evaluation design	Rolling origins with separate training, calibration and test periods
Benchmarks	Naive and seasonal-naive methods at a minimum
Accuracy	Scaled error relative to a benchmark
Calibration	Interval coverage overall and over time, with nominal level stated
Probabilistic skill	Pinball loss or another proper scoring rule
Decision value	Detection and false-alarm rates for the alerts the tool generates
Reproducibility	Software, settings and random seeds, or code where possible

Conclusion

Machine-learning cash-flow forecasting can bring sophisticated planning tools within reach of SMEs, but its value depends on how it is evaluated. In a reproducible experiment, a gradient-boosted model matched a seasonal-naive benchmark on accuracy while its uncalibrated intervals were markedly overconfident, causing more missed cash shortfalls. Conformal calibration narrowed but did not close the gap under changing conditions. On real revenue from 327 small online sellers, the accuracy ranking reversed, but the overconfidence of raw intervals and the value of conformal calibration were confirmed, and the decision-value dimension revealed an alert rule that did not work. Evaluating forecasts jointly on accuracy, calibration

and decision value gives SME owners, software providers and lenders a sounder basis for adopting AI-driven financial planning tools.

References

[1] Karim, N. B., & Seitz, J. (2023). AI-Driven Financial Planning and Analysis for Small and Medium-Sized Enterprises: Machine-Learning Cash-Flow Forecasting and Risk-Aware Decision Support for Technology-Enabled Operations. *International Journal of Modern Computing*, 6(1), 158-176.

[2] Farrell, D., & Wheat, C. (2016). *Cash is king: Flows, balances, and buffer days—Evidence from 600,000 small businesses*. JPMorgan Chase & Co. Institute.

[3] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.

[4] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437–450.

[5] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE, 13*(3), e0194889.

[6] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting, 38*(4), 1346–1364.

[7] Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 192–213.

[8] Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B, 69*(2), 243–268.

[9] Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application, 1*, 125–151.

[10] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378.

[11] Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50.

[12] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.

[13] Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.

[14] Romano, Y., Patterson, E., & Candès, E. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems* (Vol. 32).

[15] Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting, 25*(1), 3–23.

[16] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.

[17] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170.

[18] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[19] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.

[20] Olist. (2018). *Brazilian e-commerce public dataset by Olist* [Data set]. Olist. https://github.com/olist/work-at-olist-data

[21] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (pp. 3146-3154).