International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(14s) 2026 pp. 1207-1211
https://doi.org/10.70917/ijcisim-2026-4333
Article

Geometric Log-Space Ensembling for MAPE-Optimal Retail Revenue Forecasting: A Large-Scale Case Study

Daniel Dragonevskiy

Faculty of Computer Science, HSE University (National Research University Higher School of Economics), Moscow, Russia
Email: ispromashka@gmail.com
ORCID: https://orcid.org/0009-0003-0017-3454

Abstract: Forecasting monthly store-level turnover across tens of thousands of retail locations is a canonical large-scale retail analytics problem: the panel is wide and heterogeneous, and the dominant variance is multiplicative, driven by calendar effects, macro-economic cycles, and local competition. We report a forecasting pipeline built for a public retail-forecasting challenge for a grocery retailer operating more than 18,000 stores, predicting each store's revenue (PTO) for a target month from prior sales history, store metadata, and a calendar of events. Performance is evaluated with the multiplicative score $S = 100(1 - \mathrm{MAPE}/100)^2$, minimised by forecasters that are geometrically, not arithmetically, unbiased. We show analytically, and confirm via synthetic simulation, that for log-normal targets the geometric (log-space) ensemble is the MAPE-optimal aggregator of unbiased predictors. Guided by this, we (i) reparameterise the target as a multiplicative residual relative to the previous month; (ii) train CatBoost models on log-turnover with calendar features; and (iii) combine these with an independently developed external gradient-boosting model via a weighted geometric blend under a global mass-preservation constraint. This pipeline reaches a confirmed public-leaderboard score of 90.68 (MAPE ≈ 4.77%), up from a 90.28 baseline, within 0.14 of the best observed public score (90.82). We additionally report a submission-log-grounded ablation of what did not work, and discuss why common ensembling and calibration heuristics fail to transfer to this metric.
Keywords: retail forecasting, MAPE, gradient boosting, ensembling, geometric mean, mass preservation.

1. Introduction

Modern grocery retail produces some of the densest commercial time-series data in the world: every store is a multivariate, non-stationary panel whose dynamics are shaped by holidays, pricing, regional macro-economics and local competition. The retailer motivating this paper operates more than 18,000 stores that differ in their revenue dynamics across format, geography and local competitive density. This heterogeneity means that no single model captures all relevant signal with equal fidelity, which motivates a deliberately combined, multi-model pipeline rather than a single end-to-end predictor. At this scale, even a fraction of a percentage point of forecasting error translates into a material misallocation of working capital and logistics capacity.

The forecasting challenge that motivates this paper is a concrete, checkable instance of this class of problem. Given 18,657 stores and monthly revenue history over roughly two years, the task is to predict each store's turnover for a target month. Performance is measured by

$$S = 100\,(1 - \mathrm{MAPE}/100)^2, \quad \mathrm{MAPE} = (100/N)\,\Sigma\,|\hat{y}_i - y_i| \,/\, y_i, \quad (1)$$

so that a score of 90.68 corresponds to MAPE ≈ 4.77%. Because S is quadratic in (1 − MAPE/100), small reductions in MAPE translate into disproportionately large gains near the top of the leaderboard, which rewards careful, low-variance improvements over broad but noisy ones.

Our contributions are: (1) a derivation, with a supporting synthetic simulation, of why geometric (log-space) combination is the MAPE-optimal aggregation rule for log-normally distributed retail-revenue targets; (2) a practical, fully reproducible recipe — multiplicative residual reparameterisation, log-space CatBoost training with calendar features, and geometric blending under a mass-preservation constraint — validated on a real large-scale forecasting task, reaching a confirmed public leaderboard score of 90.68, within 0.14 of the best observed public score; (3) a submission-log-grounded ablation study covering more than a dozen confirmed negative results; and (4) an evidence-based discussion of a design path that did not pan out — adding neural forecasters trained from scratch as a source of ensemble diversity — and why architectural difference alone did not translate into predictive independence in this setting.

2. Related Work

Combining multiple forecasts is one of the oldest and most robust findings in the forecasting literature, going back to Bates and Granger's demonstration that a simple average of two forecasts can outperform either individually [1]. Timmermann's survey [2] catalogues why combination helps in practice and surveys weighting schemes; that literature, however, implicitly targets additive error, which is the wrong geometry when the evaluation metric is MAPE. Tofallis [3] shows that training and evaluating under MAPE implicitly assumes multiplicative errors, and that a regression trained on the log scale with a symmetric loss is the natural MAPE-consistent estimator. Kolassa [4] studies MAPE-optimal point forecasts for retail-sales distributions and notes that the geometric mean, rather than the arithmetic mean, minimises expected relative error for multiplicative-error models. The M5 competition [9] established a large public benchmark for hierarchical retail-sales forecasting and popularised gradient-boosted-tree solutions (LightGBM [7]) as strong, practical baselines for this problem class. CatBoost [8] is used throughout this work for its native handling of categorical store/region features. Neural architectures purpose-built for time series (e.g., N-BEATSx [5]) and pretrained time-series foundation models (e.g., Chronos [6]) are increasingly used as ensemble-diversity sources; Section V reports that this premise did not hold in our setting when the neural model was trained from scratch on the same internal data as the tree-based models.

3. Problem Setup and Data

The target $y_i \in \mathbb{R}_+$ is the target-month revenue (PTO) of store i, i = 1, …, N = 18,657. The empirical distribution of y_i is strongly right-skewed, consistent with the multiplicative nature of store revenue (foot traffic × basket size × price index); we treat $\log y_i$ as approximately log-normal throughout.

The organisers provide a panel of 485,082 rows (one store × month) with 24 columns: a store identifier; year and month; several point-of-sale behavioural aggregates (mean promotional items per receipt, mean basket size, mean cancellations, working hours per day); store metadata (opening-age category, trade-area-size category, city, region, population, households); local context features (pedestrian/car traffic per hour, counts of nearby marketplaces, pharmacies, schools, transit stops, competing grocery stores, and same-chain stores within 500 m); operational flags (checkout count, alcohol-licence flag); and the target revenue. One external public dataset — a regional panel of average monthly wages — was used by a collaborator's model described in Section IV-B.

We build three classes of features: (i) lagged and rolling aggregates: per-store lags at 1, 2, 3, 4, 6, 12, 13 months, and rolling means/standard deviations over 3, 6, 12 months, computed strictly on shifted (pre-target) values to avoid leakage; (ii) calendar: the weekday composition of the target month and the overlap of the target month with internationally recognised dates — specifically, whether International Women's Day (March 8) falls on a weekend, and how the Christmas week aligns with the rest of the month, both of which shift the surrounding purchasing pattern; (iii) categorical encodings: CatBoost's native ordered target statistics for region/format.

4. Method

A. Multiplicative Residual Reparameterisation

Rather than regressing revenue directly, we predict the log-ratio to the previous month, $\delta_i = \log(y_i / y_{i,\text{lag1}})$, and reconstruct the forecast as $\hat{y}_i = y_{i,\text{lag1}} \cdot \exp(\hat{\delta}_i)$. The rationale is that $y_{i,\text{lag1}}$ already explains most of the cross-sectional variance — stores change slowly month to month — so the residual model only has to learn the much lower-variance deviation from last month. On held-out validation months, this changes a same-store naive carry-forward baseline of 7.70% MAPE and a direct-log-target model at 7.80% MAPE into a 6.86–6.34% MAPE residual model, depending on the specific base learner and feature set (Table I); it is the single largest driver of predictive accuracy in the whole pipeline.

B. Base Learners

Two model families feed the final ensemble. The in-house CatBoost stack is a geometric-mean stack of four CatBoost variants trained on the residual/ratio target with different base offsets (lag_1, lag_{12}, a 3-month mean) and different losses (MAE, multi-quantile), combined with a naive per-region ratio heuristic at a 10% weight, plus a deeper CatBoost model (depth 10, 4,000 iterations, MAE loss on log revenue) trained with the calendar features of Section III. This entire stack is fully reproducible from the accompanying code.

The external gradient-boosting model is a LightGBM model with a Tweedie objective (variance power 1.2), trained by an independent collaborator on a differently reparameterised target — the ratio of revenue to a calendar-adjusted baseline — with an external regional-wage feature reported as the most informative addition to that model. We use this model's predictions as given; its training code is not part of our reproducibility package, which is a limitation discussed in Section VI.

C. Why Geometric Ensembling is MAPE-Optimal

Let $\hat{y}_1, \ldots, \hat{y}_K$ be unbiased predictors of y in the sense $E[\hat{y}_k] = y$. Their arithmetic mean $\bar{y} = \Sigma\, w_k\, \hat{y}_k$ is unbiased on the additive scale; however, MAPE error, $E[|\bar{y} - y|/y]$, is governed by the multiplicative deviation $\bar{y}/y$, not the additive deviation $\bar{y} - y$. Forming instead the geometric ensemble

$$\hat{y}^{geo} = \exp(\Sigma\, w_k \log \hat{y}_k), \quad \Sigma\, w_k = 1, \quad (2)$$

gives a predictor that is unbiased in log y, which for log-normal targets is exactly the scale on which the loss is symmetric; under log-normal errors, the geometric mean is the MAPE-optimal aggregator [4]. A controlled synthetic simulation calibrated to a log-normal distribution with parameters roughly matching the scale of store-level target-month revenue ($\mu = 11.5$, $\sigma = 0.65$) confirms this: at every blend weight, the geometric mean of two unbiased log-normal-noise predictors yields strictly lower MAPE than their arithmetic mean. On the real submission history, every attempt to replace the geometric-mean stack with an alternative aggregator — median ensembling, or per-store multiplicative recalibration — measured strictly worse on the public leaderboard (Section V).

D. Mass Preservation

Blending predictors in log space can shift the total predicted mass $M = \Sigma\, \hat{y}_i$ away from a well-calibrated anchor, even when each individual predictor is well calibrated on average. Every confirmed submission in our log that applies a global or large-subset multiplicative correction to an already-strong anchor reduces the score: scaling the baseline stack by ×1.0305 moves the score from 90.28 to 89.75 (−0.53); a smaller, bias-correction-motivated ×0.9957 rescaling moves it from 90.28 to 90.01 (−0.27); and a ×1.005 correction applied to a 1,754-store subset identified as under-predicted moves the 90.43 stack to 90.42 (−0.01). We did not find a rescaling that improved the score. The rule used in the final pipeline is therefore to preserve the anchor model's total mass exactly when forming any blend:

$$\hat{y}_i^{*} = \hat{y}_i^{geo} \cdot (M_{anchor} / \Sigma_j\, \hat{y}_j^{geo}). \quad (3)$$

E. Final Ensemble

The confirmed, submitted pipeline combines the in-house CatBoost stack with the external LightGBM-Tweedie model via a weighted geometric blend with mass preservation:

$$\hat{y}_i^{final} = M_{anchor} \cdot \exp(w\, \ell^{CB}_i + (1-w)\, \ell^{LGB}_i) / \Sigma_j \exp(w\, \ell^{CB}_j + (1-w)\, \ell^{LGB}_j), \quad (4)$$

where $\ell^{\bullet} = \log \hat{y}^{\bullet}$. Several blend weights w (including w = 0, i.e., the external model alone) scored identically at 90.68 within the leaderboard's display resolution, indicating that incorporating the in-house component via the geometric rule does not degrade the external model's performance.

5. Experiments

A. Stage-1 Ablation Ladder

TABLE I. STAGE-1 WALK-FORWARD ABLATION (4-FOLD)

Configuration	MAPE	Score S
Flat-mean baseline	45.70%	29.50

Naive carry-forward (lag_1)	7.70%	85.20
Direct log-target LightGBM (L1)	7.80%	85.01
Residual LightGBM	6.86%	86.75
Residual CatBoost	6.81%	86.84
Residual LightGBM (tuned)	6.46%	87.49
Residual CatBoost (depth 8, tuned)	6.41%	87.59
Residual CatBoost + cleaned features	6.34%	87.73
+ region/format group-aggregate features	6.65%	87.14
Direct MAPE-loss objective	14.88%	72.46
Per-store post-hoc MAPE calibration	7.88%	84.85

B. Stage-2 Leaderboard Progression

TABLE II. CONFIRMED LEADERBOARD SUBMISSIONS

Submission	Score S	Δ
Baseline: 5-seed CatBoost residual stack	90.28	—
+ variance-based routing	90.36	−0.07 (vs. 90.43)
+ routed consensus blend	90.39	−0.04 (vs. 90.43)
Median ensemble	90.40	−0.03 (vs. 90.43)
+ per-store k* multiplier	90.42	−0.01 (vs. 90.43)
In-house geometric stack	90.43	—
+ calendar-deep CatBoost (8% weight)	90.44	+0.01 (vs. 90.43)
+ external LightGBM-Tweedie model	90.67	+0.23 (vs. 90.44)
Final confirmed (mass-preserved blend)	90.68	+0.01 (vs. 90.67)
Best observed public score	90.82	—

C. What Did Not Work

We record three further confirmed negative results. Training with sample_weight = 1/y on top of an already log-transformed residual target double-counts the same correction and produces a severely biased model; the resulting submission scored 79.42, a 10.86-point drop from the 90.28 baseline — the single worst confirmed result in the project. Blending in a zero-shot foundation-model forecast (TimesFM) with an incorrectly estimated calibration multiplier (×1.044) reduced the score from 90.28 to 83.68 (−6.60). Training CatBoost with its native MAPE objective directly on the raw target produced 14.9% validation MAPE (Table I), more than double the residual-model MAPE.

D. Documented Pairwise Correlations

We additionally logged pairwise prediction correlations across candidate ensemble members. Alternative tree-based architectures and neural forecasters trained from scratch on the same internal features are all highly correlated with the in-house stack (≥ 0.98): an alternative LightGBM configuration (0.998) and a neural forecaster (N-BEATSx) trained from scratch (≈ 0.99). The only tested sources of substantially lower correlation are either a different training objective on the same tabular features (Tweedie, 0.36; hand-engineered features, 0.678) or a genuinely external prior

(pretrained foundation models, 0.85–0.92). None of the low-correlation candidates was successfully integrated into a confirmed, submitted result.

6. Limitations

The confirmed 90.68 result combines our own, fully reproducible CatBoost pipeline with an independently developed LightGBM-Tweedie model whose training code is not part of our reproducibility package — only its predictions are available to us. Our own independently confirmed, end-to-end reproducible contribution reaches 90.44. Motivated by combination theory, we explored adding a neural forecaster trained from scratch, in search of a third, genuinely independent ensemble member; this did not succeed, as correlation with the tree-based models was ≈ 0.99. Blend weights and feature choices were selected against the competition's public leaderboard under a tight submission budget, which risks a degree of overfitting to the specific public evaluation split. This is a single-competition case study on one retail network over one target month; while the underlying argument is metric-driven and should generalise to other multiplicative-error forecasting settings, the specific feature engineering is specific to the retail calendar studied here.

7. Conclusion

Under a multiplicative evaluation metric, the geometry of the loss should drive both model training and model combination. Two ingredients did the most work: reparameterising the regression target as a multiplicative residual and training on the log scale, and combining models via a weighted geometric mean in log space rather than an arithmetic mean. Mass preservation — simply not rescaling an already-well-calibrated anchor — was the cheapest and most consistently validated correction available. Genuine ensemble diversity in this problem class appears to require an external data source or an externally pretrained prior, not merely a different model architecture trained on the same internal features; pretrained time-series foundation models are a natural direction for future work, together with hierarchical reconciliation [10], [11] across the store–format–region hierarchy.

Acknowledgement

The author thanks the collaborators who contributed the external LightGBM-Tweedie model and regional wage data referenced in Section IV-B.

References

1. J. M. Bates and C. W. J. Granger, "The combination of forecasts," Journal of the Operational Research Society, vol. 20, no. 4, pp. 451–468, 1969.
2. A. Timmermann, "Forecast combinations," in Handbook of Economic Forecasting, vol. 1, Elsevier, 2006, pp. 135–196.
3. C. Tofallis, "A better measure of relative prediction accuracy for model selection and model estimation," Journal of the Operational Research Society, vol. 66, no. 8, pp. 1352–1362, 2015.
4. S. Kolassa, "Evaluating predictive count data distributions in retail sales forecasting," International Journal of Forecasting, vol. 32, no. 3, pp. 788–803, 2016.
5. K. G. Olivares, C. Challu, G. Marcjasz, R. Weron, and A. Dubrawski, "Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx," International Journal of Forecasting, vol. 39, no. 2, pp. 884–900, 2023.
6. A. F. Ansari, L. Stella, C. Turkmen, et al., "Chronos: Learning the language of time series," arXiv preprint arXiv:2403.07815, 2024.
7. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Proc. NeurIPS, 2017.
8. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in Proc. NeurIPS, 2018.
9. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," International Journal of Forecasting, vol. 38, no. 4, pp. 1346–1364, 2022.
10. R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang, "Optimal combination forecasts for hierarchical time series," Computational Statistics & Data Analysis, vol. 55, no. 9, pp. 2579–2589, 2011.
11. S. L. Wickramasuriya, G. Athanasopoulos, and R. J. Hyndman, "Optimal forecast reconciliation using a least squares approach," Journal of the American Statistical Association, vol. 114, no. 526, pp. 804–819, 2019.