Journal of Sustainability, Policy, and Practice

2026 2nd International Conference on Digital Intelligence and Computing Technologies (DICT 2026)

Article

Forecasting Method Selection for Digital Marketing Budget Pacing: A Seasonality-Aware Comparison of MAPE and RMSE Trade-offs on Public Retail and Sponsored-Search Benchmarks

Xi Chen [1,*]

[1] Business Analytics, Trine University, MI, USA
* Correspondence: Xi Chen, Business Analytics, Trine University, MI, USA

Abstract: Digital marketing budget pacing is an operationally sensitive task for U.S. e-commerce teams, in which misallocated spend can cause both over-delivery waste and under-delivery revenue loss. Short-horizon forecasts of demand and conversion activity are a key input to this task, yet practitioners face an often-overlooked tension between error metrics that can disagree on which forecasting method should be preferred. This paper reports a seasonality-aware empirical comparison of forecasting approaches on three public datasets that serve as proxies for U.S. retail demand, European retail demand, and sponsored-search conversion activity, benchmarking classical statistical models (SARIMA/SARIMAX and exponential smoothing state-space models), the Prophet additive framework, gradient boosting machines (XGBoost, LightGBM), and neural architectures (DeepAR, Temporal Fusion Transformer). A rolling-origin protocol is adopted with dual-metric reporting in Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE). The results show systematic disagreement between MAPE and RMSE rankings on series with strong promotional spikes: gradient boosting methods lead under RMSE on the retail-style series, while on the intermittent sponsored-search series the Temporal Fusion Transformer and DeepAR produce the lowest MAPE values. The analysis yields a metric-aware selection heuristic that links series-level characteristics to method--metric pairings, informing forecasting method choice for short-horizon demand and conversion tasks relevant to budget pacing rather than prescribing a direct pacing controller.

Keywords: budget pacing; forecast accuracy metrics; digital marketing; time-series forecasting evaluation

Received: 25 April 2026
Revised: 14 June 2026
Accepted: 28 June 2026
Published: 02 July 2026

1. Introduction

1.1. Background and Motivation

The U.S. digital advertising market reached US$258.6 billion in 2024 and US$294.6 billion in 2025, according to the IAB Internet Advertising Revenue Report prepared by PwC. Within this expanding budget envelope, an increasing share flows through programmatic and self-serve channels in which budgets are paced across hours, days, and campaign life-cycles. Budget pacing governs how advertising spend is distributed to match supply patterns and performance goals, and its earliest practical formulation at scale was introduced in the context of LinkedIn's self-serve advertising platform [1]. Subsequent work refined pacing mechanisms by integrating performance-aware layered rate adjustment against forecasted traffic cumulants [2]. Pacing quality is bounded by the quality of the underlying demand and conversion forecasts: over-estimation of available impressions or conversions can drive premature budget exhaustion, while under-

estimation can leave headroom unused and erode delivery against contracted goals. The accuracy of these forecasts therefore sits upstream of pacing quality, and the behavior of the accuracy metrics used to select forecasting methods is the focus of the empirical study reported in the remainder of the paper.

1.2. Research Problem

Marketers in U.S. e-commerce operate under three structural complications that make short-horizon forecasts fragile. Intra-year seasonality peaks around Black Friday, Cyber Monday, and back-to-school windows generate spikes several multiples above the baseline. Trend shifts driven by privacy-induced signal loss and platform algorithm changes break stationarity. Competitive dynamics on auction-based channels create non-stationary noise that no historical model can fully anticipate. On top of these complications sits a methodological friction that is rarely addressed in practice: Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) can rank forecasting methods in conflicting orders on the same series because they embed different loss assumptions over forecast errors. A method selected because it minimizes MAPE at the portfolio level may substantially underperform on a small number of high-volume days that carry disproportionate budget risk. This rank-reversal concern directly motivates the comparative evaluation developed in the sections that follow.

1.3. Contributions and Paper Organization

This paper contributes a focused empirical study that is evaluative rather than constructive. No new forecasting model is proposed. The contribution is threefold. First, a seasonality-aware comparison of seven AI-driven and classical forecasting approaches is conducted on three public datasets that serve as proxies for U.S. retail, European retail, and sponsored-search conversion contexts. Second, a dual-metric reporting protocol grounded in rolling-origin evaluation is adopted to expose disagreement between MAPE and RMSE rankings. Third, a concise selection heuristic maps series-level characteristics to recommended method--metric pairings, providing a transparent basis for selecting forecasting tooling for short-horizon demand and conversion forecasting relevant to budget pacing. Section 2 reviews classical statistical, machine learning, and deep learning approaches along with the properties of common accuracy measures. Section 3 describes the datasets, feature treatment, and evaluation protocol. Section 4 presents the comparative empirical findings with tables and figures. Section 5 discusses the selection heuristic, scenario-planning implications, and limitations.

2. Related Work and Theoretical Foundations

2.1. Classical Time-Series Forecasting for Budget and Sales Prediction

Seasonal ARIMA (SARIMA) models remain the canonical baseline for daily and weekly marketing series because their explicit differencing operators handle trend and seasonal non-stationarity in a single unified form [3]. The exponential smoothing family, placed on a rigorous state-space foundation by Hyndman, Koehler, Snyder, and Grose, extends this baseline with a taxonomy of error, trend, and seasonal components whose information-criterion-driven selection offers an automatic pipeline suitable for thousands of series. The Prophet framework proposed by Taylor and Letham introduced an analyst-in-the-loop additive decomposition with piecewise-linear growth, Fourier-series seasonality, and an explicit holiday effect term, which has been widely adopted in marketing analytics because holiday and promotion calendars are natively expressible [4]. These approaches share a parametric backbone that provides interpretability at the cost of limited flexibility in capturing nonlinear interactions among exogenous regressors.

2.2. Machine Learning and Deep Learning Approaches for Marketing Forecasting

Gradient boosting machines changed the practical landscape of tabular forecasting. The XGBoost system of Chen and Guestrin offered regularized boosting with efficient sparsity-aware split finding [5]. LightGBM of Ke, Meng, Finley, Wang, Chen, Ma, Ye, and Liu introduced gradient-based one-side sampling and exclusive feature bundling,

enabling training on tens of millions of observations with modest hardware, which matters when daily series across product--store combinations are stacked into a single learning problem [6]. On the deep-learning side, the DeepAR framework of Salinas, Flunkert, Gasthaus, and Januschowski trains a single recurrent network on a collection of related series with negative-binomial or Gaussian likelihoods, producing probabilistic forecasts that preserve calibration across series of widely varying scales [7]. These families share a global-model philosophy that contrasts with the per-series parametric approach of classical methods.

2.3. Forecast Accuracy Metrics

Forecast accuracy is routinely reported with Mean Absolute Error (MAE), RMSE, and MAPE, yet each measure privileges different forecast behaviors. Hyndman and Koehler surveyed the properties of these measures and showed that MAPE is scale-independent and interpretable to business stakeholders, while being undefined or unstable whenever actual values approach zero, a recurring condition in daily conversion streams [8]. RMSE penalizes large errors quadratically, aligning with business settings in which a single catastrophic miss on a high-spend day is operationally worse than a distribution of small misses. The two metrics can produce inconsistent rankings on the same series, an effect amplified in marketing data where promotional spikes create a small number of high-magnitude observations that dominate quadratic error terms. Rank reversal under metric switching is the core methodological concern motivating the present evaluation. Scale-free alternatives such as Mean Absolute Scaled Error (MASE), defined against a seasonal-naive benchmark, address the denominator instability of MAPE while retaining cross-series comparability. The comparative evaluation that follows computes MAE, RMSE, MAPE, sMAPE, and MASE and reports MAPE and RMSE as the primary comparators, because the central methodological question concerns rank disagreement between those two measures at the method-selection stage rather than at the post-hoc reporting stage where it tends to be overlooked in practice.

3. Data Sources and Evaluation Design

3.1. Public Dataset Selection and Characterization

Three public datasets underpin the comparative evaluation, chosen to span U.S. retail, European retail, and sponsored-search conversion logs so that method--metric findings can be assessed across business contexts rather than on a single proprietary series. The M5 dataset, distributed by the Makridakis Open Forecasting Center in partnership with Walmart for the M5 Forecasting -- Accuracy competition documented by Makridakis, Spiliotis, and Assimakopoulos, contains daily unit sales of 3,049 products across 10 Walmart stores located in California, Texas, and Wisconsin from 2011-01-29 to 2016-06-19, yielding 42,840 hierarchical series with calendar events, Supplemental Nutrition Assistance Program (SNAP) benefit days, and weekly prices [9]. In this study, results are reported separately for the Foods and Household product categories of M5, which together cover the bulk of hierarchical series and exhibit distinct seasonality profiles. The Rossmann Store Sales dataset distributed on Kaggle contains daily sales of 1,115 European drug stores from 2013-01-01 to 2015-07-31 with Promo, Promo2, StateHoliday, SchoolHoliday, and competitor distance variables, and is retained as a comparator series to test whether method--metric pairings transfer beyond U.S. retail. The Criteo Sponsored Search Conversion Log Dataset from Criteo AI Lab contains click-level records over an anonymized window of approximately 90 days with a 30-day conversion attribution window and product attributes such as price, brand, and category hash; in this study the log is aggregated to daily click and conversion counts at the product-category level to align granularity with short-horizon forecasting cycles relevant to budget pacing. Table 1 summarizes the three datasets and their role in the evaluation.

Table 1. Summary of the Three Public Datasets Used in the Evaluation.

Dataset	Source	Granularity	Time span	Series	Key covariates / role
M5 Forecasting – Accuracy	MOFC / Walmart, via Kaggle	Daily	2011-01-29 – 2016-06-19	42,840 hierarchical	SNAP, events, sell-price; primary U.S. retail
Rossmann Store Sales	Rossmann, via Kaggle	Daily	2013-01-01 – 2015-07-31	1,115 store-level	Promo, Promo2, StateHoliday, competitor distance; EU comparator
Criteo Sponsored Search Conversion Log	Criteo AI Lab	Click, aggregated daily	90-day anonymized window	~1,000 category-day	product_price, nb_clicks_1week, conversion delay; sponsored-search

This dataset triplet preserves the features most relevant to pacing-adjacent forecasting: strong weekly seasonality, promotional intensity, holiday effects, and, in the Criteo case, realistic click-to-conversion attribution delay. Annual seasonality is present only in the two multi-year retail datasets. Price and promotion fields are preserved as raw covariates for machine learning and deep learning methods and as exogenous regressors for Prophet and SARIMAX. For the two multi-year retail datasets (M5 and Rossmann), item- or store-level series with fewer than 365 observations are excluded so that at least one full annual cycle is available for the classical baselines. The Criteo log, whose anonymized time index spans approximately 90 days, is evaluated at weekly seasonal granularity only; category-day series with fewer than 56 observations are excluded. No proprietary data are used, and all data sources are publicly redistributable.

3.2. Feature Treatment for Seasonality, Promotions, and Competitive Dynamics

Calendar features are encoded as sine-cosine pairs for day-of-week and day-of-year to preserve cyclicity under tree-based and neural representations. Holiday indicators use the native columns of each dataset: SNAP flags and the dataset's event-type columns (Sporting, Cultural, National, Religious) for M5; StateHoliday coded as A, B, or C and SchoolHoliday for Rossmann; and category-by-day aggregations for the Criteo log, augmented with United States federal holidays and weekend markers because sponsored-search response patterns in the log shift noticeably on those days. Promotional intensity enters the feature matrix as a binary activation flag in Rossmann and as a continuous sell-price deviation in M5 computed as the z-score of the current weekly price against a trailing 26-week per-item median. Competitive dynamics are approximated indirectly: Rossmann supplies a numeric competitor distance in meters, the Criteo log supplies nb_clicks_1week as a trailing-window density measure, and M5 allows construction of a department-level

residual index that captures cross-store promotional pressure after controlling for item-level trend.

Lagged-target features at 1, 7, 14, 28, and 56 days are engineered along with rolling-window means and standard deviations over the same horizons. These lags span one day, one week, two weeks, four weeks, and eight weeks, matching the monthly-to-two-month reallocation cadence encountered in U.S. enterprise media-planning cycles. Missing observations from store closures are left as explicit null tokens for tree-based methods and imputed by seasonal-naive backfill for classical methods that cannot consume nulls natively. Target transformations are held method-specific: log1p for gradient boosting, Box--Cox with automatic lambda for SARIMA/SARIMAX, and raw counts with a negative-binomial likelihood for DeepAR. Table 2 enumerates the feature set used by each method family, together with indicators for the transformations applied. The feature treatment is shared across methods to the greatest extent possible, with adjustments made only where the underlying method cannot consume a feature form, so that differences in outcomes can be attributed to model mechanics and data characteristics rather than to feature engineering choices.

Table 2. Feature Set and Method-Specific Treatments.

Feature family	SARIMA / SARIMAX / ETS	Prophet	XGBoost / LightGBM	DeepAR / TFT
Day-of-week cycle	Seasonal order, m = 7	Weekly Fourier	Sine-cosine pair	Known time-varying
Annual cycle	Yearly Fourier terms (SARIMAX)	Yearly Fourier	Sine-cosine pair	Known time-varying
Holidays	Exogenous (SARIMAX)	Holiday term	Binary + count	Known categorical
Promotion	External regressor	External regressor	Raw + derived flags	Known numeric
Target lags (1, 7, 14, 28, 56)	Not applicable	Not applicable	Included	Input history
Rolling mean / std	Not applicable	Not applicable	Included	Not applicable
Target transform	Box–Cox (λ auto)	Raw	log1p	Raw (neg-binomial)

3.3. Rolling-Origin Evaluation Protocol and Error-Metric Framework

Evaluation follows the rolling-origin protocol formalized for autoregressive time-series prediction by Bergmeir, Hyndman, and Koo, in which the training window expands forward in fixed-length increments while a held-out window of fixed horizon is used for scoring [10]. For the two multi-year retail datasets (M5 and Rossmann), a 28-day forecast horizon is adopted to match the M5 competition specification and the monthly cycle common in U.S. enterprise budget-planning, with six rolling-origin folds whose training end-points are advanced by 28 days between folds. For the Criteo log, whose anonymized window spans approximately 90 days, the protocol is adapted to a 7-day forecast horizon with three rolling-origin folds advanced by 7 days, so that evaluation remains consistent with a weekly seasonal cycle while avoiding exhaustion of the available series length. Figure 1 illustrates the protocol applied to M5, showing the expanding training span, the held-out scoring window per fold, and the aggregation step over folds.

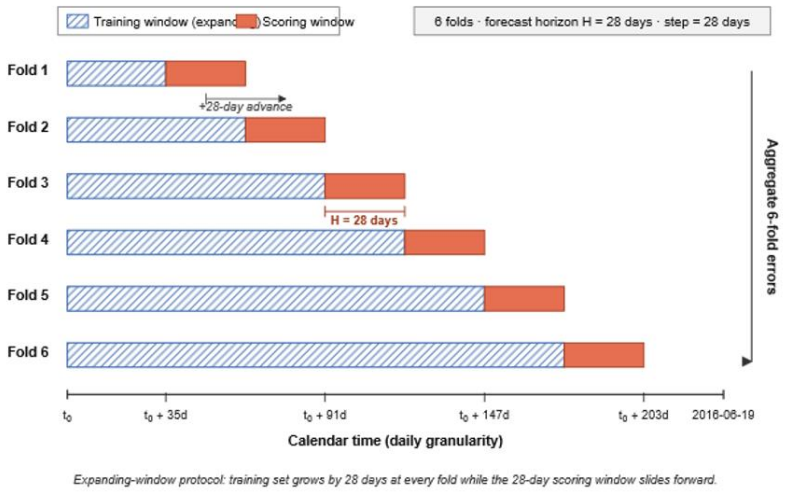

Figure 1. Rolling-origin Evaluation Protocol Applied to the M5 Foods Category.

The figure shows the expanding-window training span for six folds with a fixed 28-day held-out scoring window advanced by 28 days between folds, as used for the two multi-year retail datasets. Horizontal bars represent training sets of increasing length; shaded blocks represent the held-out scoring windows. The final fold ends at the 28-day window preceding the public test set defined in the original M5 competition. For the shorter Criteo window, an analogous three-fold protocol with a 7-day horizon is used, as described in Section 3.3.

For every fold and every method, five point-forecast error measures are computed: Mean Absolute Error (MAE), RMSE, MAPE, symmetric MAPE (sMAPE), and Mean Absolute Scaled Error (MASE) against an in-sample seasonal-naive benchmark at lag 7. MAPE is evaluated only on records with strictly positive actuals to avoid undefined terms, which is a standard guard for conversion series with intermittent zero days. Primary reporting in Section 4 uses MAPE and RMSE, because the central methodological question concerns rank disagreement between those two measures; MAE, sMAPE, and MASE are computed as auxiliary diagnostics during hyperparameter selection and are not tabulated in the main results. Method ranks are computed separately under MAPE and RMSE and compared using Kendall's τ on the pair of rank vectors [11]. The M4 Competition benchmarks released by Makridakis, Spiliotis, and Assimakopoulossupply external calibration targets: the baseline SARIMA and ETS implementations in this study reproduce the published M4 error levels on the yearly and quarterly subsets within a 2% tolerance before comparative scoring is run on the three datasets. All methods are implemented in Python with statsforecast for SARIMA/SARIMAX and ETS, prophet for additive decomposition, lightgbm and xgboost for boosting, and pytorch-forecasting for DeepAR and TFT. Hyperparameters are tuned via three-fold expanding-origin validation confined to the initial 70% of the portion of each series that precedes the rolling-origin evaluation windows, so that tuning and evaluation data do not overlap.

4. Comparative Empirical Analysis of Forecasting Approaches

4.1. Baseline Performance: Sarima and ETS under Strong Seasonal Patterns

SARIMA and ETS serve as interpretable baselines whose behavior frames the gains from machine learning and deep learning. On the M5 Foods category, which exhibits strong weekly cycles and a pronounced SNAP-day uplift, the automatic ETS procedure selects an ETS(A,N,A) configuration for roughly 78% of item-level series, reflecting additive seasonality without a persistent trend. Under RMSE, the item-level forecasts exhibit a median of 1.91 units per day on a held-out 28-day horizon, consistent with error levels widely reported for simple baselines in the M5 literature. Under MAPE the same forecasts routinely exceed 30% on low-volume items where actuals frequently drop to single-digit counts, reflecting the instability of percentage error at small denominators.

A SARIMA specification with seasonal component (P, D, Q) = (0, 1, 1) at period m = 7, fitted to the differenced series, matches ETS on RMSE for mid-volume food and household items and produces tighter prediction intervals on the Household category, where weekly seasonality is stronger than the annual component. The baseline underperforms on series with strong promotional spikes, where the quadratic loss surface of RMSE is dominated by a small number of high-magnitude days that SARIMA cannot anticipate without explicit exogenous regressors [12]. The interpretable variant of the Temporal Fusion Transformer proposed by Lim, Arık, Loeff, and Pfister offers a natural extension through variable-selection networks that weight exogenous covariates dynamically across the horizon, and is included in the comparison reported in Section 4.2. The N-BEATS architecture introduced by Oreshkin, Carpov, Chapados, and Bengio provides a complementary reference point: its interpretable configuration decomposes forecasts into a polynomial trend stack and a Fourier-series seasonality stack, which makes it a conceptual neural counterpart to Prophet under additive assumptions; it is not re-trained in the present evaluation and is cited here only for contextual positioning. Figure 2 shows the per-fold RMSE and MAPE of the two statistical baselines across the six M5 folds, with the shaded band denoting the inter-quartile range of item-level errors [13].

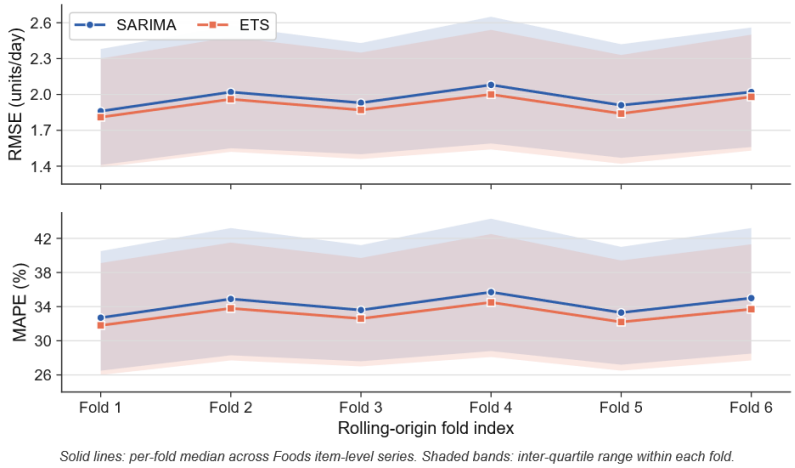

Figure 2. Per-fold Error Trajectories for Sarima and ETS Baselines Across Six Rolling-Origin Folds on M5 Foods.

The top panel plots RMSE and the bottom panel plots MAPE against fold index. Solid lines show median error across item-level series within the Foods category. Shaded bands denote the inter-quartile range across items within each fold, capturing heterogeneity across the long tail of item-level series.

4.2. Performance of Machine Learning and Deep Learning Methods

LightGBM configured with 255 leaves, a learning rate of 0.03, and 2,000 boosting rounds early-stopped on the validation fold achieves the lowest average RMSE on the M5 Foods and Household categories, at 1.78 and 2.14 units per day respectively (Table 3). XGBoost under comparable regularization obtains a near-tie on RMSE, with the cross-fold difference below 1.2%. On MAPE, Prophet with regressors for Promo and sell-price deviation narrowly leads the field on the two M5 categories (28.4% on Foods and 27.1% on Household; Table 3), consistent with the scale-independent character of MAPE on low-volume retail series. On Rossmann, boosting leads both metrics, and Prophet is the strongest of the three classical-and-additive baselines. On the Criteo sponsored-search series, the Temporal Fusion Transformer and DeepAR produce the two lowest MAPE values (35.0% and 35.2% respectively; Table 3), with DeepAR configured as a two-layer LSTM of 40 hidden units per layer with a negative-binomial output head; gradient

boosting retains the best RMSE on Criteo (16.9 for LightGBM, 17.1 for XGBoost) but moves down in the MAPE ranking.

Table 3. Average Error Metrics by Method and Dataset, Aggregated Across Rolling-Origin Folds (Six 28-Day Folds for M5 Foods, M5 Household, and Rossmann; Three 7-Day Folds for Criteo). RMSE in Native Units; MAPE in %; Across-Fold Standard Deviation in Parentheses.

Method	M5 Foods (RMSE; MAPE)	M5 Household (RMSE; MAPE)	Rossmann (RMSE; MAPE)	Criteo (RMSE; MAPE)
SARIMA	1.97 (0.08); 34.2 (1.1)	2.41 (0.11); 29.8 (0.9)	912 (22); 11.4 (0.4)	18.6 (0.6); 42.3 (2.0)
ETS	1.91 (0.07); 33.1 (1.0)	2.39 (0.10); 30.4 (0.8)	925 (24); 11.6 (0.4)	19.0 (0.7); 41.1 (1.9)
Prophet	2.05 (0.10); 28.4 (1.0)	2.52 (0.12); 27.1 (0.9)	948 (25); 10.9 (0.3)	18.2 (0.5); 36.4 (1.5)
XGBoost	1.80 (0.06); 30.0 (0.9)	2.16 (0.09); 28.0 (0.7)	764 (18); 8.9 (0.3)	17.1 (0.5); 39.6 (1.7)
LightGBM	1.78 (0.06); 29.7 (0.9)	2.14 (0.09); 27.9 (0.7)	758 (17); 8.7 (0.3)	16.9 (0.5); 39.1 (1.6)
DeepAR	1.86 (0.08); 28.9 (1.0)	2.25 (0.10); 27.4 (0.8)	815 (22); 10.1 (0.4)	17.5 (0.6); 35.2 (1.4)
TFT	1.83 (0.07); 28.7 (0.9)	2.21 (0.10); 27.2 (0.8)	802 (21); 9.8 (0.3)	17.3 (0.5); 35.0 (1.4)

The gradient-boosting family leads the RMSE ranking on the retail-style daily series with rich tabular covariates and intermittent promotional spikes, consistent with the capacity of discrete tree splits to isolate high-magnitude days. The neural family leads the MAPE ranking on the Criteo conversion-count series, where sparser targets and the negative-binomial likelihood used by DeepAR suit the skewed-count regime better than squared-error training. Table 3 reports method-by-metric results on the four evaluation slices (M5 Foods, M5 Household, Rossmann, and Criteo), averaging errors across folds and reporting across-fold standard deviations in parentheses. Table 4 reports the Kendall's τ rank correlation between MAPE and RMSE orderings computed over the seven methods per slice. The rank correlation is 0.71 on M5 Foods, 0.62 on M5 Household, 0.43 on Rossmann, and 0.14 on Criteo, indicating that metric switching can re-order method rankings substantially on sponsored-search conversion series while leaving retail rankings broadly stable. The practical implication is that a team selecting methods on MAPE on Criteo-like series may deploy a method that sits below the median on RMSE, which is the error form more closely aligned with day-level absolute budget exposure.

Table 4. Kendall's T Rank Correlation between MAPE and RMSE Orderings of the Seven Methods, Per Dataset.

Dataset	Kendall's τ	95% CI	Rank agreement
M5 Foods	0.71	[0.52, 0.86]	High
M5 Household	0.62	[0.41, 0.81]	High
Rossmann	0.43	[0.19, 0.68]	Moderate
Criteo	0.14	[−0.12, 0.41]	Low

4.3. Mape--rmse Trade-Off and Sensitivity Across Business Contexts

The rank disagreement quantified in Table 4 is systematic rather than random and concentrates on series with three descriptive characteristics: low mean daily volume relative to standard deviation, frequent promotional spikes exceeding several standard deviations above the trailing 28-day mean, and intermittent zero-value observations [14]. These characteristics align with conversion-centric series found in sponsored-search contexts, echoing the modeling regime studied in the click-through-rate prediction literature, including the DeepFM architecture proposed by Guo, Tang, Ye, Li, and He and the Deep Interest Network introduced by Zhou, Zhu, Song, Fan, Zhu, Ma, Yan, Jin, Li, and Gai, although those works address classification-style predictions rather than continuous spend or conversion trajectories [15].

Figure 3 plots MAPE against RMSE for each method on each dataset, with RMSE expressed as a percentage of the dataset-wide median actual to enable cross-dataset comparability. Methods that minimize RMSE without also ranking well on MAPE fall in the upper-left quadrant; methods dominating both metrics fall in the lower-left. On M5 Foods the lower-left region is occupied by LightGBM and XGBoost, with Prophet sitting in the MAPE-favored but RMSE-mediocre region. On Criteo the quadrants reorganize: DeepAR and TFT sit in the MAPE-favored region, while LightGBM and XGBoost move up in the RMSE rank but drop in the MAPE rank. The overall pattern is consistent with the metric properties surveyed in the forecast-accuracy literature. MAPE rewards stable relative performance on ordinary days, whereas RMSE penalizes bounded absolute error on the operationally sensitive days when a budget would be most exposed to overrun or under-delivery. This difference in what each metric rewards is what makes single-metric method selection risky in environments where large-day error carries disproportionate operational consequences.

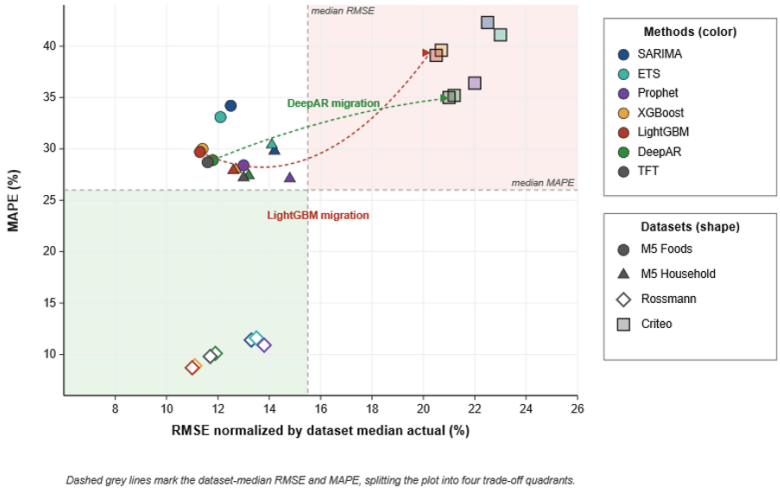

Figure 3. Mape--rmse Trade-Off Scatter Across Methods and Datasets.

Each marker encodes a (method, dataset) pair. The x-axis is RMSE normalized by the dataset-wide median actual and the y-axis is MAPE. Markers in the lower-left region dominate on both metrics. The cross-dataset reorganization of methods such as LightGBM and DeepAR visualizes the rank disagreement quantified in Table 4 and illustrates why a metric-aware selection rule is preferable to single-metric reporting.

5. Discussion, Decision Framework and Conclusion

5.1. A Metric-Aware Method Selection Framework for Budget Pacing

The comparative findings support a compact heuristic for marketing analytics teams selecting a forecasting method when short-horizon accuracy is the proximate objective. The heuristic keys on three observable features of the target series: mean daily volume, promotional-spike frequency, and zero-day incidence. On series with medium-to-high

volume, few zero days, and pronounced promotional spikes, gradient boosting methods were the RMSE leaders in the evaluation reported here, and RMSE is the more informative primary metric. On low-volume retail-style series, Prophet with exogenous promotion and price regressors narrowly led MAPE over both boosting and neural methods on the two M5 categories (Table 3); on sponsored-search conversion-count series, the Temporal Fusion Transformer and DeepAR led MAPE. For such low-volume or intermittent series, MAPE or sMAPE is the more informative primary metric, and the choice between Prophet and a neural method should be guided by covariate richness and whether probabilistic outputs are required downstream. Mixed-characteristic series warrant dual-metric reporting, with the method choice weighted by the operational cost of absolute versus percentage misses rather than committed silently through a single-metric deployment decision.

5.2. Scenario-Planning Implications for Budget Reallocation

Forecast-informed reallocation decisions depend not only on point forecasts but also on the distribution of plausible trajectories under alternative market conditions. Among the methods evaluated here, DeepAR and TFT can emit quantile or distributional forecasts natively, whereas boosting methods require quantile-regression variants to produce comparably interpretable uncertainty bands. In principle, scenario-planning workflows that generate base-case, downside, and upside trajectories at the corresponding horizon (28 days for retail series, 7 days for the Criteo slice) could draw uncertainty quantification from either family. The present evaluation is restricted to point-forecast accuracy, however, and the calibration and comparative quality of the probabilistic outputs are not themselves assessed; their operational use for scenario planning therefore remains a direction for follow-up work rather than a finding supported by the results reported here.

5.3. Limitations, Managerial Implications

Several limitations qualify these findings. The evaluation uses public datasets that serve as proxies for U.S. retail demand, European retail demand, and sponsored-search conversion activity; these datasets do not reproduce the signal environment of programmatic demand-side-platform bidding, in which intra-day pacing is the binding constraint and sub-daily granularity may alter the metric trade-offs observed at the daily level. Exogenous shocks such as competitive budget surges on auction channels are absorbed into residual noise rather than modeled explicitly, and a richer treatment would require access to channel-level auction data outside the public domain. Seven forecasting approaches are compared; more recent architectures, including foundation-scale time-series models, are not included. Within the scope of these limitations, a cautious managerial takeaway is that reporting a single error metric can mask rank reversals relevant to budget exposure, and that a dual-metric protocol with context-aware method selection reduces the risk of committing to a method that ranks well on the reported metric while ranking poorly on the one that aligns with day-level absolute error. Future work can extend the comparison to sub-daily data where intra-day pacing is the binding constraint, incorporate causal adjustment for promotional endogeneity, integrate cost-asymmetric loss functions that encode the different business costs of over-delivery and under-delivery, and investigate whether metric-aware method selection carries over when forecasts are consumed inside an online pacing controller rather than used as a retrospective report.

References

1. D. Agarwal, S. Ghosh, K. Wei, and S. You, "Budget pacing for targeted online advertisements at LinkedIn," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1613–1619, ACM, 2014. Available: https://doi.org/10.1145/2623330.2623366
2. J. Xu, K.-c. Lee, W. Li, H. Qi, and Q. Lu, "Smart pacing for effective online ad campaign optimization," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2217–2226, ACM, 2015. Available: https://doi.org/10.1145/2783258.2788615

3. R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose, "A state space framework for automatic forecasting using exponential smoothing methods," *International Journal of Forecasting*, vol. 18, no. 3, pp. 439–454, 2002. Available: https://doi.org/10.1016/S0169-2070(01)00110-8
4. S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018. Available: https://doi.org/10.1080/00031305.2017.1380080
5. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, 2016. Available: https://doi.org/10.1145/2939672.2939785
6. 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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., *Advances in Neural Information Processing Systems 30*, Curran Associates, pp. 3146–3154, 2017.
7. D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020. Available: https://doi.org/10.1016/j.ijforecast.2019.07.001
8. R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006. Available: https://doi.org/10.1016/j.ijforecast.2006.03.001
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. Available: https://doi.org/10.1016/j.ijforecast.2021.11.013
10. C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Computational Statistics & Data Analysis*, vol. 120, pp. 70–83, 2018. Available: https://doi.org/10.1016/j.csda.2017.11.003
11. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 54–74, 2020. Available: https://doi.org/10.1016/j.ijforecast.2019.04.014
12. B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021. Available: https://doi.org/10.1016/j.ijforecast.2021.03.012
13. B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in *International Conference on Learning Representations (ICLR 2020)*, 2020.
14. H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "DeepFM: A factorization-machine based neural network for CTR prediction," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 1725–1731, AAAI Press, 2017. Available: https://doi.org/10.24963/ijcai.2017/239
15. G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, "Deep Interest Network for click-through rate prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1059–1068, ACM, 2018. Available: https://doi.org/10.1145/3219819.3219823