Journal of Business, Social and Technology
Volume 7, Issue 4, 1784-1795
e_ISSN: 2807-6362
https://bustechno.polteksci.ac.id/index.php/jbt/index
DOI: doi.org/10.59261/bustechno.v7i4.755

Online Reviews in SKU-level Demand Forecasting Using Global Cross-Learning LSTM

Ivan Michael Siregar
Institut Teknologi Harapan Bangsa, Indonesia

Irfan Muhammad Fauzi*
Institut Teknologi Harapan Bangsa, Indonesia

***Corresponding author:**
Irfan Muhammad Fauzi, Institut Teknologi Harapan Bangsa, Indonesia.
✉ds-19017@students.ithb.ac.id

Article Info:
Article history:
Received: July 03, 2026
Revised: August 27, 2026
Accepted: August 30, 2026
Available Online: September 07, 2026

Keywords:
Global Cross-Learning LSTM; Intermittent Demand; Online Reviews; Proxy Theory; Sparsity Barrier.

Abstract
Background: Daily SKU-level demand forecasting is challenging in zero-inflated e-commerce data because absolute-error optimization can favor zero forecasts and fail to capture demand activation.
Objective: This study evaluates a Global Cross-Learning LSTM that combines transaction history, structural sparsity markers, and online-review signals as behavioral proxies.
Methods: Computational forecasting experiments used daily SKU-level transaction and review data from the Amazon and Olist datasets. Four ablation configurations (M0–M3) were compared with SES, Croston, and SBA using MASE, RMSSE, and AMSE. The LSTM used a 30-day lookback window, 50 units, the Adam optimizer, a batch size of 128, 20 epochs, and Huber loss. Paired differences across 22 entities were assessed using the Wilcoxon signed-rank test.
Results: Across the full timeline, M0 obtained the lowest MASE (0.4243), reflecting the advantage of zero forecasts on inactive days. During activation windows, however, M3 reduced AMSE from 0.7769 to 0.6195 and produced a statistically significant improvement in RMSSE ($p < 0.05$). Global pooling also supported forecasting for cold-start and lumpy-demand items.
Conclusion: The findings support Proxy Theory by indicating that review volume and valence can provide leading information before sufficient transaction data accumulate. Theoretically, the study links behavioral proxies with global representation learning for intermittent demand. Practically, the model can inform replenishment and safety-stock decisions by reducing underforecasting during demand activation. The novelty lies in jointly evaluating online-review proxies and global cross-learning at the SKU level under lifecycle-specific conditions.

To cite this article: Siregar, I. M., & Fauzi, I. M. (2026). Online Reviews in SKU-level Demand Forecasting Using Global Cross-Learning LSTM. *Journal of Business, Social and Technology*, *7*(4), 1784-1795. https://doi.org/10.59261/bustechno.v7i4.755

INTRODUCTION

Forecasting daily SKU-level demand in modern e-commerce is challenging because demand arrivals are intermittent and nonzero demand sizes are erratic. At this level, the available data exhibit a pronounced sparsity barrier: 91.69% of the examined SKUs were classified as intermittent (Caserta & D'Angelo, 2025; Kourentzes & Athanasopoulos, 2021). SES, Croston's method, and the Syntetos-Boylan Approximation (SBA) address intermittency through univariate historical patterns, but they do not incorporate behavioral signals that may precede transactions.

In zero-inflated series, absolute-error measures such as MASE can favor zero forecasts because absolute error is minimized by the median. Such forecasts may score well during inactive periods yet underforecast demand activation, with direct consequences for replenishment and stockout exposure. Series-by-series machine-learning models also tend to overfit short, zero-heavy histories (Makridakis et al., 2018; Spiliotis et al., 2022). Thus, the central problem is not only model accuracy but also whether the evaluation design identifies forecasts that respond adequately to demand spikes.

Demand forecasting at the granular SKU level is inherently challenged by two structural factors: variable demand arrivals, known as intermittency, and irregular demand sizes, known as erraticness. In the forecasting literature, these dimensions are formally characterized using two specific metrics: the Average Inter-Demand Interval (ADI), which measures the average interval between successive nonzero demand occurrences, and the Squared Coefficient of Variation (CV^2), which captures the relative variability of nonzero demand sizes. Together, high intermittency and high demand-size variability characterize demand as “lumpy” or highly irregular (Kourentzes & Athanasopoulos, 2021; Pattnaik et al., 2026).

The combination of intermittency and erraticness creates a severe “sparsity barrier” for forecasting models. Traditional statistical models, such as Croston’s method and the Syntetos-Boylan Approximation (SBA), attempt to address this challenge by separately estimating demand size and inter-demand intervals (Spiliotis et al., 2022). However, they remain constrained by their reliance on univariate historical data and may struggle to capture exogenous, nonlinear market shocks. Furthermore, applying standard machine-learning (ML) algorithms directly to highly erratic and intermittent items immediately encounters this sparsity barrier. When ML models are trained in a traditional “series-by-series” fashion on bottom-level data, they may overfit or fail to converge because zero-heavy historical sequences provide insufficient continuous signals for effective learning (Lolli et al., 2017; Zhang et al., 2024).

For zero-inflated demand, MASE can favor conservative zero forecasts because absolute error is minimized by the median. RMSSE instead places greater emphasis on larger demand spikes, while AMSE captures systematic forecast bias. These complementary measures are therefore useful for distinguishing overall point-forecast accuracy from responsiveness during demand activation (Kolassa, 2016; Kourentzes & Athanasopoulos, 2021; Pattnaik et al., 2026).

Squared-error and bias measures are particularly relevant for inventory decisions because they expose errors that absolute-error summaries can obscure. This study therefore reports MASE, RMSSE, and AMSE together rather than treating any single measure as definitive (Kolassa, 2016; Kourentzes & Athanasopoulos, 2021; Pattnaik et al., 2026).

To bypass the sparsity barrier, supply-chain practitioners have historically resorted to data aggregation, grouping data across temporal or cross-sectional dimensions to smooth variance and create continuous, learnable macro-signals. While temporal aggregation can successfully reduce intermittency, it inherently reduces the granular SKU visibility required for precise, item-level inventory control (Athanasopoulos et al., 2024).

More recently, cross-learning has emerged as a methodological alternative to aggregation. Rather than modeling each series in isolation, cross-learning models extract information from multiple time series simultaneously. For daily SKU demand, training a global deep-learning network concurrently across thousands of items allows the model to identify shared latent patterns and leverage statistical strength across series (Smyl, 2020; G. Yang et al., 2024). In highly intermittent environments, this global pooling can act as an implicit regularizer, reducing the overfitting that can plague traditional series-by-series machine learning on short, zero-heavy sequences. This enables the network to generalize demand patterns across sparse items while preserving granular SKU-level visibility (Di Fonzo & Girolimetto, 2023).

While cross-learning addresses the structural limitations of sparse data, overcoming the limited transactional information requires integrating exogenous behavioral signals. In digital marketplaces, social engagement can act as a powerful behavioral proxy, shaping pre-purchase expectations and reducing product-fit uncertainty (Nugroho & Wang, 2024).

Crucially, the behavioral impact of these social proxies is particularly pronounced during specific phases of the product lifecycle, especially during activation windows when an item transitions from zero to nonzero sales. Empirical evidence suggests that the initial accumulation of product reviews acts as a substantial behavioral signal, with the first five reviews serving as leading indicators associated with increases in consumer purchase likelihood of up to 270% (Center, 2017; Jain & Singh, 2026). Furthermore, the valence of this engagement may have an asymmetric impact on demand shocks. Research demonstrates that relatively rare one-star reviews can carry disproportionate weight with consumers compared with five-star reviews, driving significantly stronger downward adjustments in sales expectations (Sahoo et al., 2018; Wang et al., 2026). By dynamically integrating multidimensional behavioral proxies, such as sentiment valence and social-proof intensity, deep-learning networks may anticipate underlying behavioral shifts that precede platform-level demand spikes, effectively using engagement as a proxy for demand before transactional histories accumulate (Anoop & Rahman, 2025; Li et al., 2020).

Table 1
Comparison of Intermittent Demand Forecasting Paradigms

Forecasting Paradigm	Core Mechanism	Handling of the Sparsity Barrier	Information Scope	Granularity
Traditional Statistical (SES, Croston, SBA)	Univariate smoothing of history	Isolates demand sizes from inter-demand intervals	Information vacuum (relies strictly on past sales)	SKU-Level
Standard Machine Learning	Series-by-series training	Fails/Overfits (lacks sufficient continuous signals)	Univariate history + limited structural features	SKU-Level
Top-Down / Aggregation	Macro-level hierarchical forecasting	Bypasses sparsity entirely by grouping data	Continuous macro-signals	Category-Level (Destroys SKU visibility)
Global Cross-Learning LSTM (Proposed)	Global concurrent pooling across items	Acts as automatic regularizer by sharing statistical strength	Multi-dimensional (Integrates behavioral proxies & social proof)	SKU-Level

Previous studies have demonstrated the potential of global and machine-learning approaches for demand forecasting, but important methodological gaps remain. Smyl (2020) combines exponential smoothing with recurrent neural networks to exploit information shared across time series, whereas Spiliotis et al. (2022) compare statistical and machine-learning methods for daily SKU forecasting. Yang et al. (2024) further demonstrate the applicability of LSTM-based modeling to e-commerce behavior.

Taken together, the literature reveals three interrelated gaps. First, although global cross-learning can mitigate the sparsity barrier by pooling information across SKUs, existing research has not jointly incorporated structural intermittency characteristics (ADI and CV^2) and online-review behavioral signals into a global SKU-level forecasting architecture. Second, although online reviews may provide behavioral information preceding transactions, their role as leading proxies for demand activation under sparse and intermittent conditions remains insufficiently established. Third, conventional forecasting evaluations do not explicitly distinguish overall point-forecast accuracy from responsiveness to demand activation, limiting their ability to assess the inventory relevance of forecasts for cold-start, newly activated, and lumpy-demand products.

This study addresses these gaps through a Global Cross-Learning LSTM that combines historical demand with ADI, CV^2, and online-review information. The study makes three main contributions. First, it contributes a multimodal global forecasting framework for intermittent SKU-level demand. By integrating structural demand characteristics with behavioral proxies, the proposed framework moves beyond the traditional reliance on historical transactions and examines whether review volume and sentiment can provide incremental information about demand before sufficient sales history has accumulated.

Second, it contributes a lifecycle-specific evaluation framework for intermittent demand forecasting. Rather than evaluating models solely on aggregate forecast accuracy, the study distinguishes full-timeline performance from performance during zero-to-nonzero demand activation. MASE, RMSSE, and AMSE are jointly employed to capture complementary dimensions of forecast accuracy, spike sensitivity, and systematic bias across relevant SKU lifecycle cohorts.

Third, it contributes empirical evidence on the incremental value of structural and behavioral information. Using M0–M3 ablation configurations, the study systematically isolates the contribution of historical demand, structural sparsity indicators, and online-review behavioral proxies. This enables the study to determine whether these additional information sources improve forecasting performance and whether their value varies across intermittent, cold-start, newly activated, and lumpy-demand SKUs.

The novelty of the study therefore lies not simply in applying an LSTM to intermittent demand, but in connecting global representation learning with structural sparsity characteristics and observable consumer-engagement signals within a lifecycle-specific forecasting and evaluation framework. Theoretically, this provides an empirical test of whether online behavioral proxies can provide leading information about demand activation before transaction histories become sufficiently informative. Practically, it provides an inventory-oriented basis for identifying forecasts that are not only accurate on average but also sufficiently responsive to emerging demand, thereby supporting replenishment decisions and reducing the risks associated with underforecasting demand activation.

METHOD

This study adopted a quantitative computational-experimental and predictive-modeling design. A comparative ablation framework was used to evaluate whether online reviews provided useful behavioral proxies when SKU-level sales histories were sparse. The stages comprised data alignment, quality screening, sparsity classification, feature engineering, normalization, sequence construction, pooled Global Cross-Learning LSTM training, benchmark forecasting, lifecycle-cohort evaluation, and paired statistical testing.

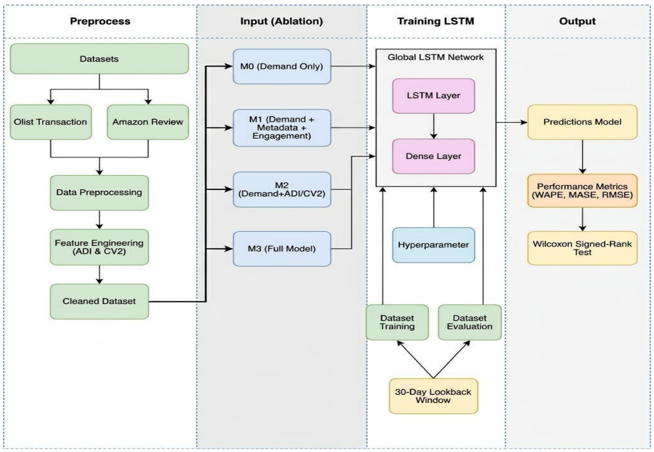

Figure 1. Framework of SKU-level Deman Forcasting

Dataset and SKU-level Sparsity Classification

The empirical evaluation used daily SKU-level transaction and online-review data from the Amazon and Olist e-commerce datasets. The analysis retained the granular, zero-inflated series and evaluated a targeted cohort of 50 highly volatile SKUs. The available experiment record did not report separate marketplace-level SKU counts, observation counts, or calendar coverage; these values were therefore neither inferred nor fabricated.

To formally quantify the structural volatility of the data, the Average Inter-Demand Interval (ADI) and Squared Coefficient of Variation (CV2) were calculated for each SKU. The formulas for these structural metrics were defined as follows:

$$ADI = \frac{\sum_{i=1}^{N} ti}{N} ADI = \frac{\sum_{i=1}^{N} ti}{N} \qquad (1)$$

$$CV^2 = \left(\frac{\sigma}{\mu}\right)^2 CV^2 = \left(\frac{\sigma}{\mu}\right)^2 \qquad (2)$$

In which, ti represents the interval between demand occurrences, N is the total number of nonzero demand periods, σ is the standard deviation of the demand size, and μ is the mean demand size. Analysis of the dataset confirmed a severe sparsity barrier: up to 91.69% of the SKUs exhibited strictly intermittent sales histories (ADI ≥ 1.32 and CV2 < 0.49), while 8.31% exhibited highly lumpy demand (ADI ≥ 1.32 and CV2 ≥ 0.49). A targeted cohort of 50 highly volatile SKUs was isolated for the final ablation and performance evaluation.

Feature Engineering: Structural Markers and Behavioral Proxies

Feature engineering comprised two parallel components:

1) Structural layer: ADI and CV2 provided explicit indicators of each SKU's sparsity and erraticness (Caserta & D'Angelo, 2025; Kourentzes & Athanasopoulos, 2021).
2) Behavioral-proxy layer: review volume, average star rating (review valence), and helpfulness votes represented observable consumer engagement during periods with limited sales information (Caserta & D'Angelo, 2025; Kourentzes & Athanasopoulos, 2021; Li et al., 2020).

Before training, all features and target-demand values were mapped to the interval [0, 1] using Min-Max scaling. Dynamic variables (demand, price, and review measures) were scaled within each SKU, whereas static structural variables (ADI and CV2) were scaled globally to retain between-SKU differences and avoid division-by-zero errors. The preprocessing workflow aligned daily records, removed exact duplicate records, retained zero-demand observations as substantive intermittency rather than treating them as missing values, and limited scaling parameters to the training data to prevent leakage. No unreported outlier deletion was assumed because demand spikes were analytically meaningful.

Global Cross-Learning LSTM Architecture

Traditional machine learning and statistical models are typically trained in a "series-by-series" fashion on individual items, which frequently leads to overfitting or a failure to converge on short, intermittent sequences (Lolli et al., 2017; Spiliotis et al., 2022). To address this limitation, this study proposed a Global Cross-Learning LSTM architecture.

The pooled network used a 30-day sliding lookback represented as three-dimensional tensors. Its compact architecture contained one LSTM layer with 50 units and ReLU activation, followed by a Dense output layer. The common representation was learned across products, allowing sparse items to borrow statistical strength from patterns observed elsewhere while preserving SKU-level predictions (Yang et al., 2024; Y. Yang, 2025).

The model was trained for 20 epochs using Adam, with a batch size of 128 and Huber loss. Each SKU contributed at most 20 randomly selected sequences to reduce gradient dominance by high-volume items. Huber loss was used consistently throughout the study because its quadratic response to small errors and linear response to large errors limited the influence of extreme spikes. The original experiment record did not specify a custom learning rate, dropout rate, or early-stopping rule; these values were therefore not invented in this revision.

Ablation Configurations and Lifecycle Cohorts

To mathematically isolate the predictive impact of online reviews, an ablation study was conducted. The network was trained and evaluated under four distinct configurations:

1) M0 (Univariate Baseline): Used historical demand only.
2) M1 (Metadata + Reviews): Used demand, price, subcategory, average rating, helpfulness, and verified-purchase indicators, without ADI or CV2.
3) M2 (Univariate + Structural): Used demand, ADI, and CV2, without online-review variables.
4) M3 (Full Multivariate Model): Used demand, ADI, CV2, and the complete set of online-review variables.

The performance of these deep learning configurations was benchmarked against traditional statistical models implemented as rolling one-step-ahead forecasts. For Simple Exponential Smoothing (SES), the smoothing parameters were dynamically optimized on the training history. For the intermittent-specific baselines, Croston's method and the Syntetos-Boylan Approximation (SBA), a smoothing factor of $\alpha = 0.1$ was used for both demand size and demand interval, with the SBA applying the standard $1 - \alpha/2 = 0.95$ correction factor to mitigate bias.

Evaluation Metrics and Bias Assessment

Standard error metrics can be highly misleading in intermittent-demand environments. Because the median of sparse datasets is often zero, absolute error measures can mathematically reward models that conservatively predict flatlines of zero, systematically failing to capture demand spikes (Kolassa, 2016; Pattnaik et al., 2026). Therefore, this study evaluated performance using three scaled metrics to capture point accuracy, squared magnitude, and bias:

Mean Absolute Scaled Error (MASE)

MASE evaluates the absolute point accuracy of forecasts. Because it relies on absolute errors, MASE is mathematically minimized by predicting the median of the demand distribution. When zero observations dominate the data, optimizing for absolute error often leads a model to conservatively predict zero to minimize MASE.

$$MASE = \frac{1}{h}\sum_{t=n+1}^{n+h}\frac{|y_t-\hat{y}_t|}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t-y_{t-1}} \qquad (3)$$

Root Mean Squared Scaled Error (RMSSE)

Unlike MASE, RMSSE evaluates squared error magnitude. Squared-error metrics are minimized by predicting the mean of the demand distribution rather than the median. Targeting the mean is essential in intermittent-demand settings to ensure that the model responds to demand surges rather than strictly defaulting to zero.

$$RMSSE = \sqrt{\frac{\frac{1}{h}\sum_{t=n-1}^{n+h}(y_t-\hat{y}_t)^2}{\frac{1}{n-1}\sum_{t-2}^{n}(y_t-\hat{y}_{t-1})^2}} \qquad (4)$$

Absolute Mean Scaled Error (AMSE)

AMSE evaluates systematic forecasting bias, such as chronic under-forecasting or over-forecasting tendencies. Instead of taking the absolute value of every individual prediction error, it first sums the raw errors, allowing positive and negative errors to cancel each other out, and then takes the absolute value of the total sum. This metric is critical in inventory control for evaluating whether a model suffers from dangerous under-forecasting bias.

$$AMSE = \frac{1}{h}\,\frac{\left|\sum_{t=n+1}^{n+h}(y_t-\hat{y}_t)\right|}{\frac{1}{n-1}\sum_{t-2}^{n}|y_t-y_{t-1}} AMSE = \frac{1}{h}\,\frac{\left|\sum_{t=n+1}^{n+h}(y_t-\hat{y}_t)\right|}{\frac{1}{n-1}\sum_{t-2}^{n}|y_t-y_{t-1}} \qquad (5)$$

Learning curves evaluated training fractions ranging from 50% to 100%. The final performance evaluation considered a targeted cohort of 50 highly volatile SKUs, while the available paired statistical output contained 22 entities for the Wilcoxon signed-rank comparison. Accordingly, the Wilcoxon test was reported only for these 22 paired entities and should not be interpreted as representing statistical testing across all 50 SKUs. The nonparametric Wilcoxon signed-rank test was used because intermittent-demand error distributions are highly skewed and therefore did not justify reliance on normality assumptions. The available statistical output confirmed $p < 0.05$ for the reported activation comparison; exact Z statistics and effect sizes were not available in the supplied experiment record and were therefore not reported.

RESULTS AND DISCUSSION

The ablation results are reported separately for the full timeline, activation windows, cold-start items, and lumpy-demand items. This separation distinguishes performance during extended inactive periods from performance at operationally important transitions to nonzero demand.

Full-Timeline Performance and Metric Interpretation

Across the full timeline, M0 achieved the lowest MASE (0.4243), whereas M3 recorded 0.6159. This pattern indicates that a demand-only model can minimize absolute error by predicting zero on many inactive days. The higher MASE of M3 should therefore be interpreted alongside RMSSE, AMSE, cohort results, and forecast overlays rather than as isolated evidence of inferior forecasting.

Table 2

Ablation Study Performance Metrics

Scope	Model	MASE	RMSSE	AMSE	MASE change vs M0	RMSSE change vs M0	AMSE change vs M0
Full	M0	0.4243	0.4520	0.1311	+0.00%	+0.00%	+0.00%
Full	M1	0.4932	0.4521	0.1366	-16.24%	-0.02%	-4.20%
Full	M2	0.5891	0.4570	0.1932	-38.84%	-1.11%	-47.37%
Full	M3	0.6159	0.4578	0.1993	-45.16%	-1.28%	-52.02%
Cold Start	M0	0.4843	0.4684	0.1243	+0.00%	+0.00%	+0.00%
Cold Start	M1	0.5687	0.4689	0.1568	-17.43%	-0.11%	-26.15%
Cold Start	M2	0.6616	0.4720	0.2252	-36.61%	-0.77%	-81.17%
Cold Start	M3	0.6797	0.4722	0.2319	-40.35%	-0.81%	-86.56%
Activation	M0	1.1619	0.9014	0.7769	+0.00%	+0.00%	+0.00%
Activation	M1	1.2105	0.8986	0.7210	-4.18%	+0.31%	+7.20%
Activation	M2	1.3007	0.8954	0.6373	-11.95%	+0.67%	+17.97%
Activation	M3	1.3222	0.8953	0.6195	-13.80%	+0.68%	+20.26%

The metric differences arise from their loss geometry. MASE weights each absolute deviation linearly and is minimized by the median, which is often zero in sparse series. RMSSE squares deviations and consequently assigns greater influence to missed demand spikes. AMSE aggregates signed errors before taking the absolute value, making persistent under- or overforecasting visible. The observed ranking changes are thus consistent with differences in sensitivity to inactive periods, spike magnitude, and directional bias (Kolassa, 2016; Pattnaik et al., 2026).

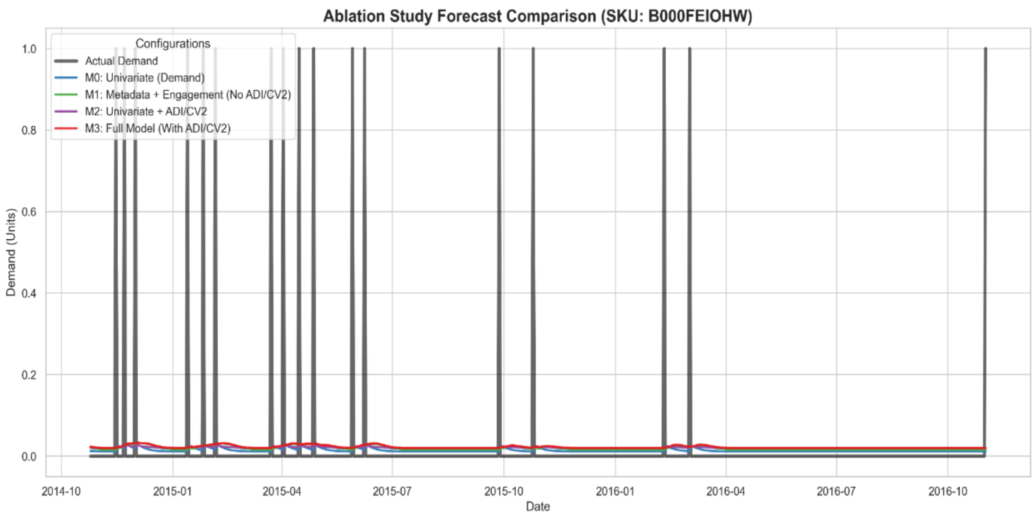

Figure 2. Overlay prediction against models

For inventory management, a low full-timeline MASE can coexist with insufficient responsiveness to demand activation. RMSSE and AMSE therefore complement MASE by indicating the magnitude of missed demand spikes and the direction of systematic error, both of which affect replenishment quantities and safety-stock decisions.

Performance During Activation Windows and Proxy Theory

Activation windows are the transitions from zero to nonzero sales, when accumulated transaction history is least informative and an inventory decision is most sensitive to a missed demand spike. This lifecycle interval provides a direct test of whether online reviews contribute information before sales are observed.

During activation windows, M3 reduced AMSE from 0.7769 for M0 to 0.6195, corresponding to a 20.26% reduction in forecasting bias relative to the demand-only baseline. This improvement reflects the combined contribution of structural sparsity indicators and online-review variables included in the full multivariate specification. More specifically, M2, which incorporates demand, ADI, and CV2 without online-review variables, achieved an AMSE of 0.6373, whereas M3 further reduced AMSE to 0.6195. This additional reduction of approximately 2.79% indicates that online-review information provides incremental predictive value beyond historical demand and structural sparsity indicators during demand activation. Nevertheless, M3 does not minimize full-timeline MASE, indicating a trade-off between overall absolute-error performance and responsiveness during operationally important activation periods.

A Wilcoxon signed-rank test assessed whether paired improvements were consistent across the evaluated SKUs rather than being driven by a small number of SKUs. Figure 3 presents the distribution of paired differences.

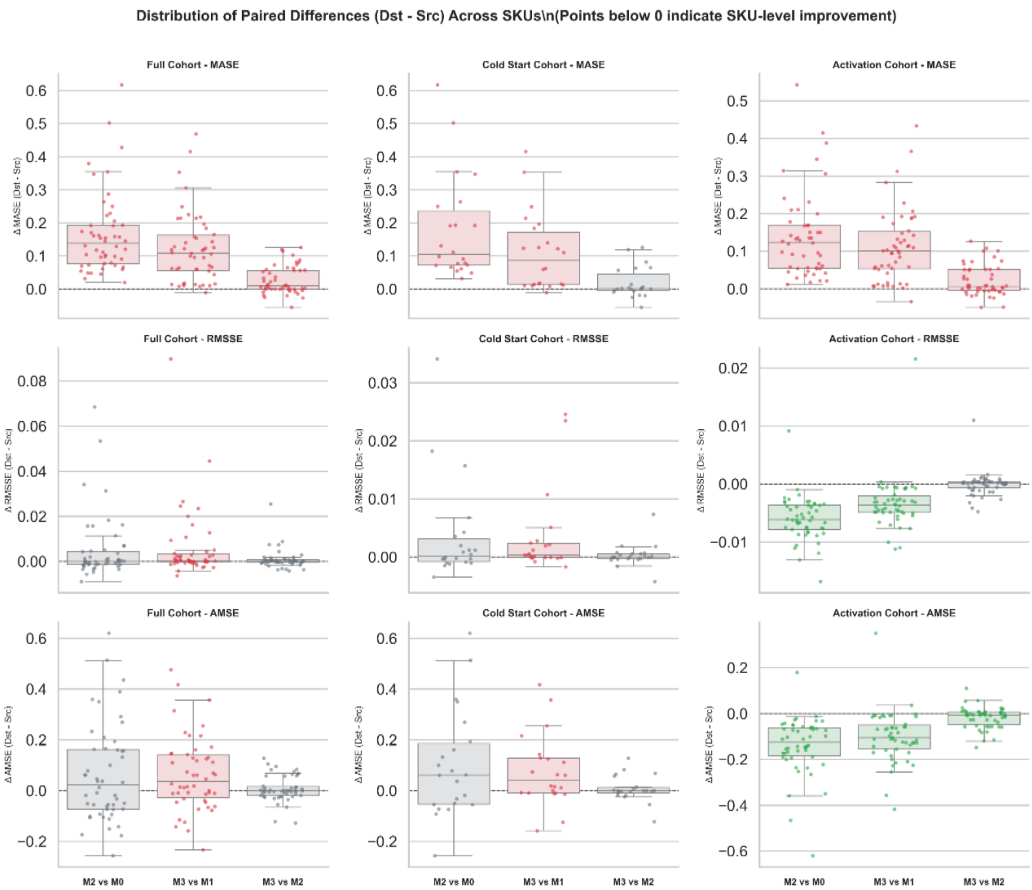

Figure 3. Distribution of Paired Performance Differences

The activation-cohort panels show a concentration of M3-versus-M1 differences below zero for squared error and bias. The Wilcoxon test indicates a statistically significant difference in RMSSE during activation ($p < 0.05$). Exact Z statistics and effect sizes were unavailable in the supplied analysis output; therefore, this revision reports only the verified significance level.

These results support, rather than conclusively prove, Proxy Theory: online-review activity may provide leading information before sufficient transaction histories have accumulated. This interpretation is consistent with evidence that review signals affect consumer expectations and e-commerce behavior (Wang et al., 2026; G. Yang et al., 2024).

Cross-Learning for Lumpy Demand and Cold-Start Items

Global pooling learns recurrent representations from heterogeneous SKUs and constrains sparse items to leverage patterns supported elsewhere in the portfolio. This shared representation reduces reliance on the limited observations available for a single cold-start or lumpy-demand series and functions as a form of regularization, while SKU-specific input sequences retain item-level context.

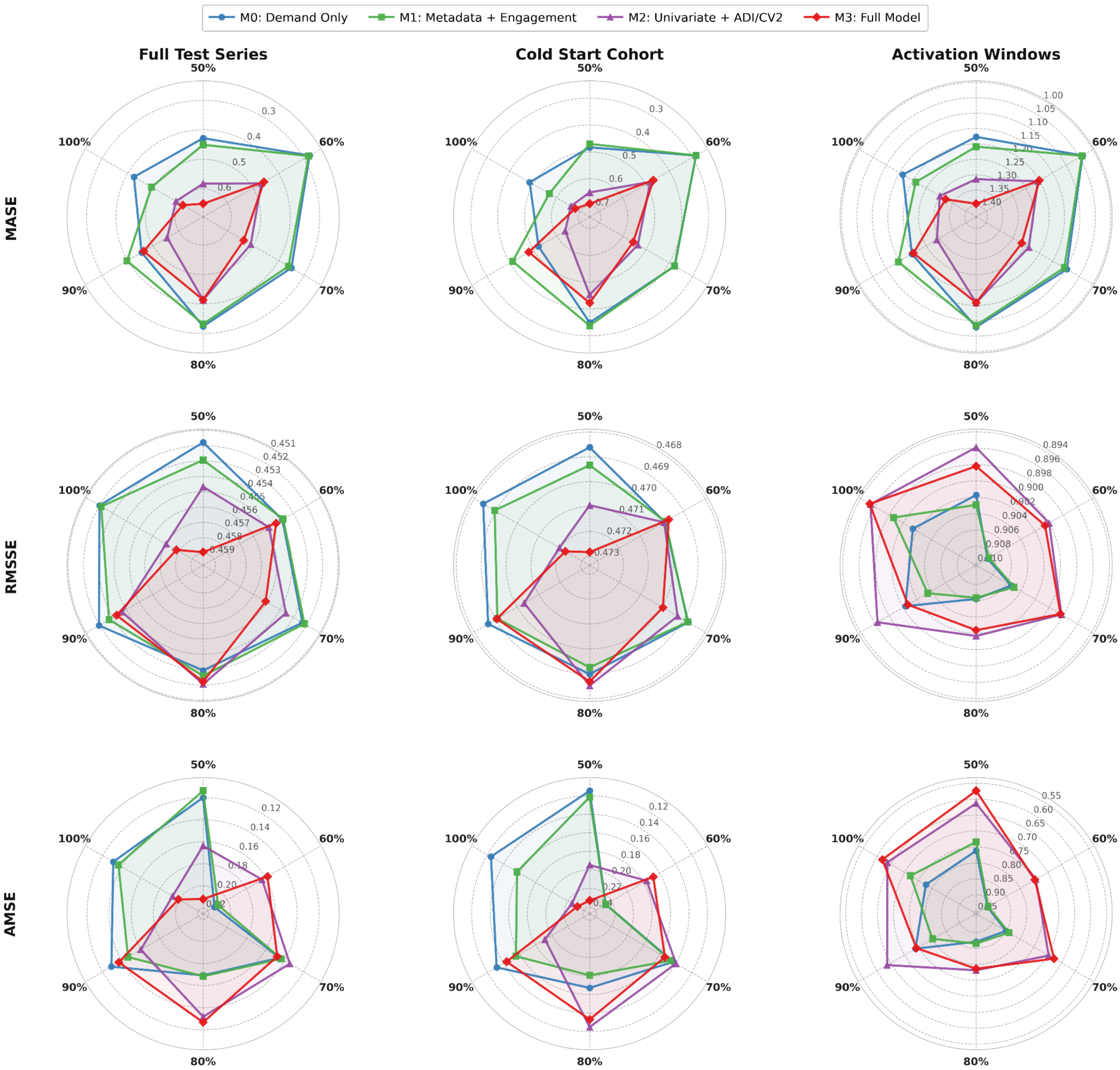

Figure 4. Ablation learning curve against all scopes

The cohort results identify two practical settings in which this representational mechanism is relevant:

1) Cold-start SKUs: For products with fewer than 98 historical reviews, individual histories are too sparse for stable series-specific estimation. Relative to SBA, the Global Cross-Learning LSTM reduced MASE by 16.32% and RMSSE by 6.69%, indicating that shared representations can improve extrapolation when local evidence is limited.
2) Lumpy demand: For items combining intermittent arrivals with highly variable positive demand (CV2 ≥ 0.49), pooled representations expose the network to a broader range of spike patterns. This mechanism can improve the learned response to erratic variation, whereas univariate smoothing estimates depend primarily on each item's limited history.

Taken together, the findings favor a lifecycle-specific evaluation strategy: full-timeline MASE describes average absolute accuracy, whereas activation-window RMSSE and AMSE reveal responsiveness and bias at operationally consequential transitions.

Discussion

The results extend earlier forecasting research in three ways. Consistent with Makridakis et al. (2018) and Spiliotis et al. (2022), they show that metric choice and sparse training histories materially affect model assessment. In line with Smyl (2020) and Yang et al. (2024), global learning provides a mechanism for sharing information across related series. The present study adds a lifecycle-specific test of online reviews as behavioral proxies within that global architecture.

Interpretation of Absolute and Squared Error Metrics

M0's full-timeline MASE of 0.4243 reflects strong absolute-error performance across numerous inactive observations, but it does not, by itself, establish an adequate response to demand spikes. This result is consistent with Kolassa (2016) and Kourentzes and Athanasopoulos (2021): absolute-error, squared-error, and bias measures address different forecasting questions (Pattnaik et al., 2026). Accordingly, inventory-oriented evaluation should interpret MASE together with RMSSE, AMSE, and lifecycle-cohort evidence.

Online Reviews as Behavioral Proxies

The activation-window results provide conditional evidence for the incremental value of online reviews as behavioral proxies. M2, which incorporates historical demand and structural sparsity indicators without online-review variables, achieved an AMSE of 0.6373, whereas M3, which additionally incorporates the complete set of online-review variables, reduced AMSE to 0.6195. This approximately 2.79% incremental reduction suggests that review-related information contributes additional predictive information beyond demand history, ADI, and CV2 during zero-to-nonzero demand transitions. Unlike Smyl (2020), which combines recurrent learning with statistical forecasting, and Yang et al. (2024), which models e-commerce behavior end to end, the M0–M3 ablation framework enables the contributions of behavioral and structural information to be examined separately. These findings provide conditional support for Proxy Theory within the evaluated datasets and lifecycle cohorts rather than establishing a universal causal effect of online reviews on demand.

Representation Learning for Lumpy Demand and Cold-Start Items

The cold-start RMSSE reduction relative to SBA indicates that the Global Cross-Learning LSTM can transfer portfolio-level temporal representations to items with limited local histories. This mechanism complements the comparative findings of Spiliotis et al. (2022) and the global-learning rationale of Smyl (2020): shared parameters regularize estimation across heterogeneous series, while review and structural inputs preserve item-specific context. The evidence supports a robust alternative to isolated models but does not imply universal superiority across domains.

Study Limitations: The evidence is limited to two e-commerce datasets and the reported SKU cohorts. The behavioral layer uses review metadata rather than text-based sentiment embeddings, multimodal product information, or graph relationships. Exact marketplace-specific observation counts, calendar coverage, Wilcoxon Z statistics, and effect sizes were unavailable in the supplied experiment record. The findings should therefore not be generalized beyond comparable e-commerce settings without external validation.

CONCLUSION

This study demonstrates that intermittent SKU-level demand forecasting benefits from integrating global cross-learning, structural demand characteristics, and behavioral signals, while evaluating performance under lifecycle-specific conditions. First, the results show that incorporating ADI, CV^2, and online-review variables into the Global Cross-Learning LSTM can improve responsiveness to demand activation. Compared with M0, the full multivariate model (M3) reduced activation-window AMSE from 0.7769 to 0.6195 and produced a statistically significant RMSSE difference ($p < 0.05$), supporting the value of combining structural and behavioral information. Second, the findings demonstrate the importance of lifecycle-specific and complementary evaluation. Although M0 achieved the lowest full-timeline MASE (0.4243), M3 performed better during demand activation. This indicates that overall accuracy alone may not adequately capture forecast responsiveness. MASE, RMSSE, and AMSE should therefore be considered jointly to assess overall accuracy, spike sensitivity, and forecast bias. Third, the M0–M3 ablation results provide conditional support for Proxy Theory, indicating that online-review variables contribute incremental predictive information beyond historical demand and structural sparsity indicators during demand activation. Practically, the Global Cross-Learning LSTM can support replenishment and safety-stock decisions for newly activated, cold-start, and lumpy-demand SKUs. Future research should validate these findings across broader datasets and examine advanced architectures, sentiment embeddings, and multimodal product information.

Future research should validate the proposed framework across broader e-commerce datasets and product categories to assess its generalizability. Further studies could also explore advanced architectures, such as Time Series Transformers and Temporal Fusion Transformers, together with richer review-sentiment embeddings and multimodal product information. Finally, future work should examine whether these behavioral signals remain predictive across different product lifecycle stages and market conditions.

ACKNOWLEDGEMENT

The authors acknowledge the providers of the publicly available Amazon and Olist datasets and the reviewers whose comments strengthened the manuscript. No additional funding source is reported in the available manuscript record.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed to the conceptualization, methodology, analysis, manuscript preparation, and critical revision of the study. All authors approved the final manuscript. Individual author names and role allocations should follow the journal submission metadata.

REFERENCES

Anoop, T. S., & Rahman, Z. (2025). Online impulse buying: a systematic review of 25 years of research using meta regression. *Journal of Consumer Behaviour*, *24*(1), 363–391.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Panagiotelis, A. (2024). Forecast reconciliation: A review. *International Journal of Forecasting*, *40*(2), 430–456. https://doi.org/10.1016/j.ijforecast.2023.10.010

Caserta, M., & D'Angelo, L. (2025). Intermittent demand forecasting for spare parts with little historical information. *Journal of the Operational Research Society*, *76*(2), 294–309.

Center, S. R. (2017). *How Online Reviews Influence Sales: Evidence of the Power of Online Reviews to Shape Customer Behavior*. Northwestern University, Medill Spiegel Research Center. https://spiegel.medill.northwestern.edu/how-online-reviews-influence-sales/

Di Fonzo, T., & Girolimetto, D. (2023). Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives. *International Journal of Forecasting*, *39*(1), 39–57. https://doi.org/10.1016/j.ijforecast.2021.08.004

Jain, A., & Singh, S. (2026). Long tail patterns in online reviews for search and experienced goods: Evidence from Amazon. *Acta Psychologica*, *269*, 107408.

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, *32*(3), 788–803. https://doi.org/10.1016/j.ijforecast.2015.12.004

Kourentzes, N., & Athanasopoulos, G. (2021). Elucidate structure in intermittent demand series. *European Journal of Operational Research*, *288*(1), 141–152. https://doi.org/10.1016/j.ejor.2020.05.046

Li, K., Chen, Y., & Zhang, L. (2020). Exploring the influence of online reviews and motivating factors on sales: A meta-analytic study and the moderating role of product category. *Journal of Retailing and Consumer Services*, *55*, 102107. https://doi.org/10.1016/j.jretconser.2020.102107

Lolli, F., Gamberini, R., Regattieri, A., Balugani, E., Gatos, T., & Gucci, S. (2017). Single-hidden layer neural networks for forecasting intermittent demand. *International Journal of Production Economics*, *183*, 116–128. https://doi.org/10.1016/j.ijpe.2016.10.021

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning

forecasting methods: Concerns and ways forward. *PLOS ONE*, *13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889
Nugroho, A., & Wang, W.-T. (2024). Credibility of YouTube product reviews and customers' product return intention: A product fit uncertainty perspective. *Information Technology & People*, *37*(4), 1560–1586. https://doi.org/10.1108/ITP-11-2021-0838
Pattnaik, M., Kumar Padhi, S., Panda, L., Naushad, U., Behera, R. K., Mishra, A., & Pattnaik, A. (2026). AI-driven sales forecasting in retail: insights from Big Bazaar's sales data. *International Journal of Management Science and Engineering Management*, *21*(1), 1–14.
Sahoo, N., Dellarocas, C., & Srinivasan, S. (2018). The impact of online product reviews on product returns. *Information Systems Research*, *29*(3), 723–738. https://doi.org/10.1287/isre.2017.0736
Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75–85. https://doi.org/10.1016/j.ijforecast.2019.03.017
Spiliotis, E., Makridakis, S., Semenoglou, A.-A., & Assimakopoulos, V. (2022). Comparison of statistical and machine learning methods for daily SKU demand forecasting. *Operational Research*, *22*(3), 3037–3061. https://doi.org/10.1007/s12351-020-00605-2
Wang, Y., Yu, B., & Chen, J. (2026). Effects of product online reviews on product returns: A review and classification of the literature. *International Transactions in Operational Research*, *33*(1), 143–176. https://doi.org/10.1111/itor.13502
Yang, G., Li, K., Zhang, Y., & Feng, Q. (2024). End-to-End Modeling and Long Short-Term Memory Application in Time Series Modeling: Modeling and Prediction of Electronic Commerce User Behavior. *Journal of Organizational and End User Computing*, *36*(1), 1–27. https://doi.org/10.4018/JOEUC.349732
Yang, Y. (2025). Data-driven prediction of future purchase behavior in cross-border e-commerce using sequence modeling with PSO-tuned LSTM. *Plos One*, *20*(12), e0337932.
Zhang, G. P., Xia, Y., & Xie, M. (2024). Intermittent demand forecasting with transformer neural networks. *Annals of Operations Research*, *339*(1), 1051–1072.