International Business & Economics Studies
ISSN 2640-9852 (Print) ISSN 2640-9860 (Online)
Vol. 8, No. 2, 2026
www.scholink.org/ojs/index.php/ibes

Original Paper

Forecasting Intermittent Demand for Luxury Goods Considering Zero-Inflated and Exogenous Shocks

Wangxin Zuo[1]

[1] Business School, University of Shanghai for Science and Technology, Shanghai, China

Received: May 8, 2026 Accepted: June 4, 2026 Online Published: June 8, 2026

doi:10.22158/ibes.v8n2p251 URL: http://dx.doi.org/10.22158/ibes.v8n2p251

Abstract

Traditional forecasting models often struggle with the intermittent demand, long-tail distribution, and zero-inflation inherent in luxury goods sales. To address this, we propose a three-stage hybrid ensemble forecasting model, RF-XGBoost-ZG. First, a Random Forest regressor captures primary time-series trends. Second, an XGBoost regressor fits the logarithmic residuals to capture nonlinear demand shocks from exogenous events. Finally, a Zero-value Gate classifier applies hard truncation and probabilistic smoothing to correct invalid demand. Empirical results demonstrate that the model significantly reduces extreme errors, balancing peak prediction accuracy with long-tail noise reduction. This framework provides a robust methodological tool for luxury retail inventory optimization and supply chain management.

Keywords

Luxury sales forecasting, Intermittent demand, Hybrid ensemble model, Residual correction, Zero-value gate

1. Introduction

Luxury retail encounters supply chain bottlenecks in the inventory management at the style level. Different from mass fashion, it has the characteristics of high price, short cycle and low tolerance for out-of-stock; excessive inventory will occupy funds, while out-of-stock will indeed damage the brand reputation. Highly accurate prediction is of great significance, but the extreme zero-inflation, sparsity and long-tail distribution make the conventional models really ineffective. To fill this gap a three - stage hybrid model (RF - XGBoost - ZG) is developed. RF captures basic lifecycle trends, XGBoost models nonlinear errors through residual correction, and then a zero - value gating module filters false demands through classification to solve the regression noise issue. An empirical analysis is conducted on the real luxury goods sales data, and it is confirmed that this model possesses the abilities of improving accuracy

and identifying zero demand.

2. Related Work

2.1 Traditional Statistical Methods

Although machine learning is becoming more and more widespread, traditional statistical methods still remain the key standards for evaluating predictive performance as Markridakis and others hold the view. Through the empirical analysis of the M5 competition (Makridakis, Spiliotis, & Assimakopoulos, 2022), the basic situation of the stationary sequence statistical method is determined and it is pointed out at the same time that there are limitations in capturing complex nonlinear patterns. Particularly in the case of intermittent demands, Koulenzeres and Teck (2021) point out that although traditional methods such as Croston can correct zero - value deviations, their generalization abilities in extremely sparse data environments are often not as good as those of neural network models.

2.2 Machine Learning and Deep Learning Models

In order to solve the problems of non - linearity and long - sequence dependence, deep - learning models emerge rapidly. The Informer model put forward by Zhou et al. emerges. With the probability sparse attention mechanism, the long sequence prediction that goes beyond the calculation bottleneck of the traditional Transformer is realized. In order to deal with the pain point of poor model interpretability, the TFT model that Lim et al. (2021) developed has successfully quantified the specific contributions of covariates like promotions and prices to sales volumes. In the fashion retail domain, Lorente - Leyva et al. (2021) further came up with the asymmetric loss function, which indicates that integrating business constraints directly into the deep learning training process can indeed greatly enhance the economic efficiency of prediction.

2.3 Ensemble Models and Residual Correction Strategies

Owing to the fact that a single model has difficulty in dealing with trends and extreme situations, the ensemble model makes use of their respective complementary strengths as Lin et al. (2021). Combining ARIMA-LSTM with variational mode decomposition to reduce prediction variance is what (Chen, Wang, Zhou et al., 2022) mentions. For zero-inflated data, Li et al. (2024) Verifies a two - stage classification - regression framework that separates demand probabilities and quantities. Moreover, There is also Samar and the others (Samal & Ghosh, 2026). The residual learning is integrated into the ensemble and the tree model is employed to correct the residuals of deep learning. This architecture, while optimizing the peak accuracy through residual correction, also retains the ability to capture trends, thereby laying a foundation for our model.

Short life cycles along with zero inflation and high volatility lead to rather scarce luxury retail frameworks. Conventional models that provide smoothed-out predictions do miss extreme bimodal distributions. The present integrated model focuses on peak errors yet neglects the end truncation, leading to a situation where inventory backlog is accelerating. Therefore, it is quite essential to integrate the peak residual correction and the lifecycle gating mechanism into a certain aggregate entity.

3. Model Architecture and Methodology

3.1 Model Architecture Overview

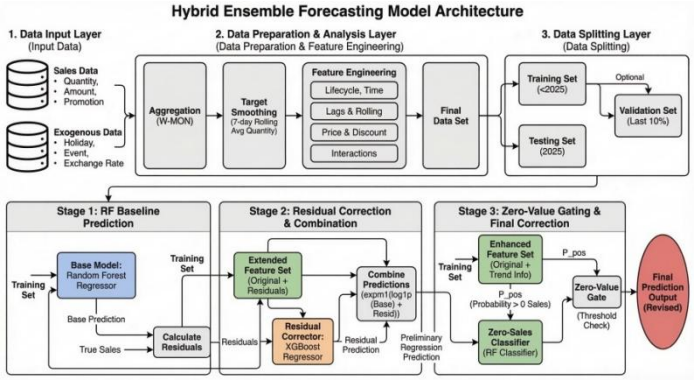

Figure 1. Architecture of the "RF-XGBoost-ZG" Model

3.2 Random Forest Module

Random forest (RF) algorithm, which is an ensemble learning method based on the bagging strategy, is constructed by building multiple independent decision trees and then aggregating their respective prediction results through regression or classification in order to obtain the final output. Its main advantage lies in having both the samples and the feature perturbations incorporated, which effectively reduces the variance of the model and makes the model possess a fairly strong resistance to overfitting. Figure 2 illustrates the architectural framework of the RF model. A random forest obtains preliminary benchmark predictions, y_{base}, by integrating the outputs of independently grown decision trees, $h_k(x)$:

$$\hat{y}_{base} = \frac{1}{K}\sum_{k=1}^{K} h_k(x) \quad (1)$$

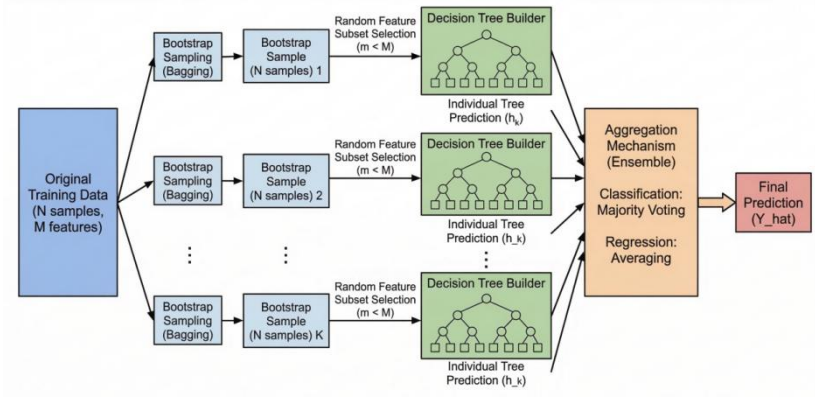

Figure 2. Structure of the Random Forest Model

3.3 Residual Correction Module

Based on the enhancement paradigm, the residual correction makes use of a residual model to approximate the errors of the base model instead of the target variable. Aggregating the two models can

capture the unaccountable patterns like nonlinear fluctuations and outliers. To address the long-tail distribution characteristics of sales volume and ensure the non-negativity constraint, the model calculates the residual between the actual sales volume and the baseline forecast value in the log space:

$$r = \log(1+) - \log(1+\hat{y}_{base}) \quad (2)$$

3.4 XGBoost Module

From the perspective of the enhancement framework, XGBoost (extreme gradient boosting), which is an efficient gradient boosting decision tree algorithm, is different from random forest. It uses an additive model to iteratively train new trees in order to fit the prediction residuals of the previous model. By incorporating the second-order Taylor expansion and the regularization term into its objective function, the model complexity and fitting accuracy of XGBoost are effectively balanced indeed. The structure of the model is presented in Figure 3.

XGBoost optimizes for log-residual errors. In round t, its objective function is approximated via a second-order Taylor expansion as:

$$O^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (3)$$

(Here, and represent the first-order gradient and the second-order Hessian matrix of the loss function, respectively, and denotes the tree regularization term used to control model complexity.)

Let be the residual prediction from XGBoost. By applying the inverse exponential transformation and adding it back to the baseline prediction, we obtain the preliminary regression prediction, y_{reg}:

$$\hat{y}_{reg} = \exp\left(\log(1+\hat{y}_{base}) + \hat{r}\right) - 1 \quad (4)$$

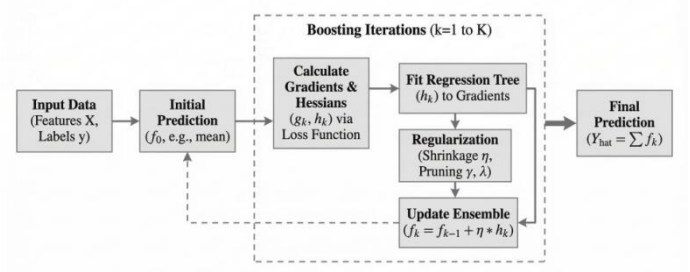

Figure 3. XGBoost Model Architecture

3.5 Zero-valued Gating Module

Two-stage zero-inflated models manage excessive zeros. A Random Forest binary classifier acts as a gate: regression predictions are kept if its probability exceeds a threshold, else zeroed. This isolates non-zero signals in actuarial science, fault detection, and sparse demand forecasting. Let's consider a random forest classifier to predict the probability of a sample generating effective demand (y>0). Given that the total number of classification trees is M, and the predicted class for a single tree is $c_m(x) \in \{0,1\}$, the probability is defined as:

$$p = P(y>0 \mid x) = \frac{1}{M} \sum_{m=1}^{M} I(c_m(x)=1) \quad (5)$$

(Among these, $I(\cdot)$ denotes the indicator function.)

The model seeks the optimal classification threshold for various business scenarios by minimizing the comprehensive business loss function τ^*, which integrates weighted errors from key metrics such as RMSE and WMAPE:

$$\tau^* = \arg\min_{\tau \in (0,1)} L(\tau) \tag{6}$$

The final predicted value is hard-truncated and smoothed based on the relationship between probability and the threshold τ^*:

$$\hat{y}_{final} = \begin{cases} \hat{y}_{reg} \cdot p^{\alpha}, & \text{if } p \geq \tau^* \\ 0, & \text{if } p < \tau^* \end{cases} \tag{7}$$

(Here, is the scene smoothing factor. When the probability falls below a threshold, dead stock noise is forcibly filtered out; when the probability meets the requirement, the predicted value is smoothed and scaled using p^{α}.)

4. Data Feature Analysis and Feature Engineering

4.1 Data Sources and Preprocessing

This research has made use of the entire-channel sales data from 2023 to 2025 of a top-ranking luxury brand in China, and this data includes the directly-owned stores and the official e-commerce platforms. This data set has 572290 transaction records that correspond to 392 different SKUs.The details of the variables are presented in Table 1.

Table 1. List of Variables in the Dataset

Variable Name	Type	Remark
Sales Date	Date	Date of Sale
Season	Categorical	Product launch season
Style Name	Categorical	Product style and design
Color	Categorical	Color
Ctry of origin	Categorical	Country of Origin
Quantity	Numerical	Sales volume
RRP	Numerical	Recommended Retail Price
Net Amount(USD)	Numerical	Final transaction amount after order discounts (excl. taxes, shipping & indirect fees)
Promotion Type	Numerical	Promotion type of the brand
First date of Sales	Date	Date of the first sale of the product
Last date of Sales	Date	Date of the st sale of the product

The preprocessing has eliminated the negative returns and the test anomalies and filled in the missing received amounts. The text dates have been transformed into date-times, and the missing sales end dates have been complemented with the latest specific-style records. The holidays, promotional activities and exchange rates are sorted out according to dates so as to integrate various external factors. Finally, the daily data is aggregated on a weekly basis for the purpose of reducing noise.

4.2 Data Feature Analysis

Figure 4 shows that the sales of luxury goods present a right - skewed long - tail distribution where the top ten styles account for 80% of the total sales and the rest often have zero records, with non - zero records only making up 13%. Well, 54% of the samples show that there is a very serious situation of data scarcity, so it is indeed necessary to develop a highly robust prediction model. Additionally, Figure 5 shows the non-linear demand surge situation brought about by holidays and promotional events such as “Double Eleven” and “Double Twelve”, while the traditional time-series frameworks are not able to capture such a situation.

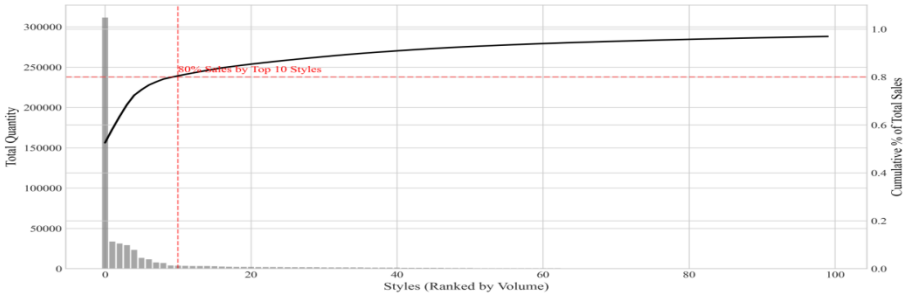

Figure 4. The Long-tail Distribution of Sales for the Luxury Brand

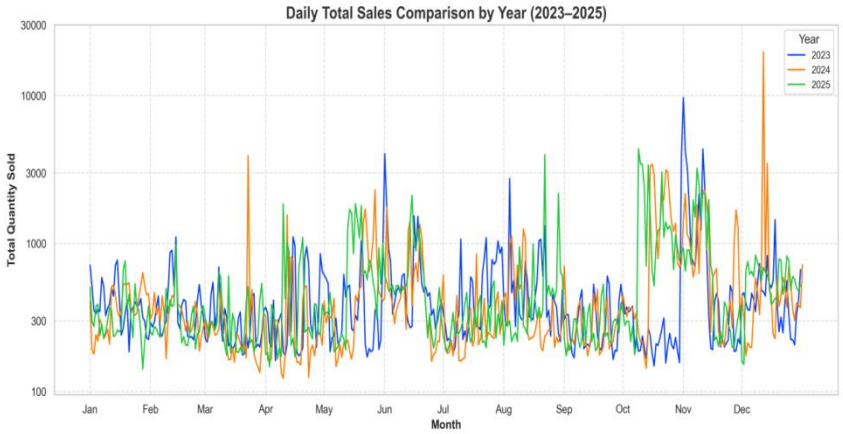

Figure 5. Daily Total Sales for the 3 Years Included in the Data

4.3 Feature Engineering

We developed a four-dimensional feature set. Temporal features include short-term lags from one to four weeks, long-term lags of 8, 12, 26, and 52 weeks, and rolling window metrics covering the mean, standard deviation, and maximum over 4, 8, and 12 weeks. Pricing features comprise a discount rate, lagged discount rates, and a price ratio comparing the current price to the four-week average. Calendar and event features utilize cyclical week-of-year encoding along with one-hot encoded holidays and promotions. Lifecycle features include weeks since launch and cumulative sales volume to track demand transitions across product stages.

5. Experimental Setup and Results Analysis

5.1 Experimental Setup

We employ forward rolling window validation with 2023-2024 training data and 2025 test data. Out-of-bag estimation mitigates overfitting and data leakage in random forest; logarithmic transformation of XGBoost residuals enhances small-batch stability and ensures non-negative outputs. Grid search optimizes module-specific hyperparameters, with the best RMSE-based configuration shown in Table 2.

Table 2. Results of Model Hyperparameter Optimisation

Model	Hyperparameter	Description	Search Space	Optimal Value
RF	n_estimators	Number of trees	[100, 200, 500]	500
RF	min_samples_split	Minimum samples for splitting	[2, 5, 10]	5
RF	min_samples_leaf	Minimum samples at leaf node	[1, 2, 4]	2
XGBoost	n_estimators	Number of boosting rounds	[100, 600, 800]	800
XGBoost	max_depth	Maximum tree depth	[3, 4, 5, 6]	5
XGBoost	learning_rate	Learning rate	[0.01, 0.02, 0.03, 0.05]	0.03
XGBoost	reg_alpha	L1 regularization coefficient	[0, 0.1, 0.3, 1.0]	0.3
XGBoost	subsample	Subsample ratio	[0.8, 0.85, 0.9]	0.8
Zero-valued Gating	n_estimators	Number of trees	[0.95, 0.995, 0.999]	0.995
Zero-valued Gating	threshold	Decision threshold	[100, 200]	200

Due to zero-inflation and long-tailed luxury data, we evaluate the model using RMSE, MAE, MAPE (y>0) for active demand where y exceeds zero, F1-score, and MASE. RMSE captures peak sales volumes by penalizing large errors, whereas MAE measures average deviations. MAPE (y>0) for active demand evaluates non-zero accuracy to eliminate zero-sales bias. The F1-score quantifies zero-value gating performance under data sparsity, and MASE determines overall model effectiveness.

5.2 Experimental Results and Analysis

Standard random forest and standard extreme gradient boosting are employed as benchmarks to verify the prediction performance, while a gating - free hybrid ensemble offers an ablation baseline to verify the zero - gating module. The experimental results are as shown in Table 3.

Table 3. Experimental Results

Model	RMSE	MAE	MAPE (y>0)	F1-Score	MASE
Standard RF	158.56	7.80	76.87%	0.529	0.985
Standard XGB	118.74	5.47	52.39%	0.812	0.812
Ensemble Model(No-Gating)	118.74	2.11	29.95%	0.886	0.812
Ensemble Model	53.04	2.05	27.82	0.953	0.389

The proposed model outperforms all baseline and ablation benchmarks across all metrics. Its RMSE drops to 53.04, a 55.3% reduction from Standard XGB and 66.5% from Standard RF, proving that residual correction successfully captures global trends and local fluctuations. Furthermore, the gating module elevates the F1-score from 0.886 to 0.953 by filtering regression noise on zero-value instances, optimizing classification accuracy, MAE, and MASE. Figures 6 and 7 compare predictions for the top three best-selling and bottom three long-tail sparse items to evaluate the Ensemble model's predictive robustness under extreme demand divergence.

In Figure 6, the standard random forest and standard extreme gradient boosting (XGB) underestimate the promotion peaks, whereas the non-gated ensemble overestimates them. In Figure 7, the standard random forest underfits the long-tail items, and the standard extreme gradient boosting and non-gated ensemble generate false peaks and noises during zero-sales intervals. Conversely the complete RF - XGBoost - ZG model can achieve the optimal fitting in all predictions because its gating mechanism can eliminate those false fluctuations during the zero - sales period and can also capture the non - linear extreme promotion situations as well as the sparse demand peaks. The RF - XGBoost - ZG model that can grasp the high - frequency volatility of the best - selling products and also resist the zero - inflation noise in the sparse long - tail products can overcome the limitations of the traditional machine learning in the intermittent demand forecasting and can support the lean inventory as well as the flexible luxury goods supply chains.

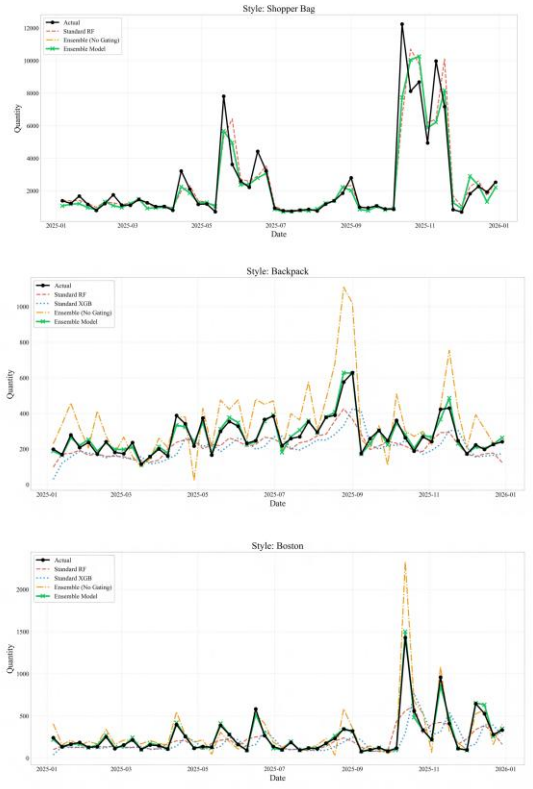

Figure 6. Comparison of Actual Sales and Predicted Results for Top 3 Bestselling Headwear Styles

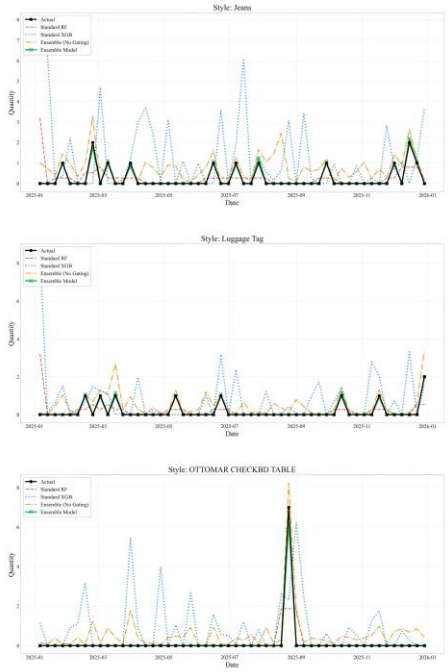

Figure 7. Comparison of Actual Sales and Predicted Results for Bottom-tier Sparse Models (Bottom 3)

6. Discussion and Conclusion

Empirical results show that the Zero-valued Gating (ZG) mechanism effectively handles sparse, zero-inflated data. Conventional models may overfit or generate false peaks on zeros, but ZG uses a binary probability threshold to filter non-demand noise. When combined with XGBoost residual correction, it captures extreme promotional peaks and improves predictive accuracy and robustness. Managerially, the framework offers crucial decision support for luxury retail with short lifecycles, high prices, and long-tail distributions, as forecasting errors can lead to stockouts or capital tie - ups. It identifies slow - moving styles to issue early markdown or outlet transfer warnings, optimizing cash flow. Meanwhile, it detects explosive peak demand to guide inventory pre - positioning, minimizing missed sales and reducing costs. This study validates a three-stage predictive model integrating Random Forest (RF), XGBoost, and Zero-valued Gating (ZG) for luxury style prediction. Empirical testing proves it overcomes traditional method limitations under long-tail distributions and volatile exogenous events. Mitigating stockout costs and inventory overhang from style proliferation, the model accurately flags tail-end lifecycles and promotional peaks to guide initial ordering, inter-store transfers, and markdown clearance.

RF-XGBoost-ZG model limitations include independent style forecasting and manual feature reliance. Future work could integrate self-attention or graph neural networks with tree structures to capture inter-style dependencies. Additionally, the single-brand dataset limits scope; future studies should use larger cross-brand and cross-regional data to validate generalization.

References

Chen, Y., Wang, Y., Zhou, Y. et al. (2022). Short term power load forecasting based on BES-VMD and CNN-Bi-LSTM method with error correction. *Frontiers in Energy Research*, *10*. https://doi.org/10.3389/fenrg.2022.1076529

Li, Z., Huang, M., & Xu, X. (2024). Tree-based machine learning intermittent demand forecasting for spare parts in electric vehicle manufacturing. *World Electric Vehicle Journal*, *15*(3), 127.

Lim, B., Arik, S. O., Loeff, N. et al. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(4), 1748-1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Lin, S. W., & Chen, K. C. (2021). On the application of ARIMA and LSTM to predict order demand based on short lead time and on-time delivery requirements. *Processes*, *9*(7). https://doi.org/10.3390/pr9071157

Lorente-Leyva, L. L., Alemany, M. M. E., Peluffo-Ordóñez, D. H. et al. (2021). Demand forecasting for textile products using statistical analysis and machine learning algorithms. *International Conference on Applied Technologies* (pp. 182-195). Cham: Springer. https://doi.org/10.1007/978-3-030-73280-6_15

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 accuracy competition: Results,

findings and conclusions. *International Journal of Forecasting*, *38*(4), 1346-1364. https://doi.org/10.1016/j.ijforecast.2021.11.013

Samal, T., & Ghosh, A. (2026). Ensemble-based predictive analytics for demand forecasting in multi-channel retailing. *Expert Systems with Applications*, *299*. https://doi.org/10.1016/j.eswa.2025.130212

Turk, B., Gök, A. A., Taştan, O. et al. (2021). Forecasting intermittent and sparse time series: A unified probabilistic framework via deep renewal processes. *PLOS One*, *16*(11). https://doi.org/10.1371/journal.pone.0259764

Zhao, N., Li, W., Wang, X. et al. (2025). Research on e-commerce inventory sales forecasting model based on ARIMA and LSTM algorithm. *Mathematics*, *13*(11). https://doi.org/10.3390/math13111838

Zhou, H., Zhang, S., Peng, J. et al. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 11106-11115. https://doi.org/10.1609/aaai.v35i12.17325