Issues in Information Systems
Volume 27, Issue 4, pp. 722-731, 2026

DOI: https://doi.org/10.48009/4_iis_2026_161

Comparative Evaluation of XGBoost, LSTM, and Benchmark Models for Retail Demand Forecasting

Hesham Allam, *Ball State University, hesham.allam@bsu.edu*
Edward J. Lazaros, *Ball State University, ejlazaros@bsu.edu*
Allen D. Truell, *Ball State University*, *atruell@bsu.edu*
Fred Kitchens, *Ball State University*, *fkitchens@bsu.edu*

Abstract

In retail supply chains, good demand forecasting is necessary for better inventory management and operational efficiency. In this paper, we compare the forecasting performance of Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) networks on the publicly available Store Item Demand Forecasting dataset, which contains 913,000 daily sales observations across 10 stores and 50 products. To preserve the temporal structure of the forecasting problem, we used a chronological train-test split. The model performance was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). The results indicate that XGBoost outperformed the LSTM model, with MAE 6.09, RMSE 7.91, and R2 0.937, compared to MAE 9.16, RMSE 10.88, and R2 -0.089. The feature importance analysis shows that lag-based variables and rolling demand statistics are the best predictors of future sales. The results indicate that XGBoost provides a more accurate and reliable approach than LSTM for retail demand forecasting using structured sales data, enabling more efficient inventory planning and supply chain decisions.

Keywords: demand forecasting, XGBoost, LSTM, retail operations, machine learning, comparative analysis

Introduction

Demand forecasting is an essential part of retail operations. Inaccurate demand forecasting can negatively affect the efficiency and profitability of retail operations (Babai et al., 2022). Traditionally, retailers have used statistical models such as exponential smoothing and Autoregressive Integrated Moving Average (ARIMA) to forecast demand (Lei et al., 2024). These models are still popular today because they are easier to understand and require fewer computations. Nevertheless, demand in retail operations is typically nonlinear due to factors such as promotions and seasonality. Statistical models may not be sufficient for handling nonlinear data (Makridakis et al., 2020).

With increasing data volumes and the ability to process large volumes of information, organizations are increasingly turning to machine learning models as an alternative to conventional models for demand forecasting. Machine learning models can better capture nonlinear patterns compared to other models (Spiliotis et al., 2022).

Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) networks are among the models that have gained significant attention for demand forecasting in retail operations (Saha et al., 2024). XGBoost is a tree-based ensemble model that uses gradient boosting and regularization techniques to generate predictive models (Chen & Guestrin, 2016). This model is commonly used for structured data problems because it is highly accurate and computationally efficient. LSTM networks are a type of recurrent neural network that can effectively handle sequential data and long-term dependencies (Hochreiter & Schmidhuber, 1997).

However, how these models perform on retail data remains poorly understood, as deep learning models require large amounts of data to outperform other models such as XGBoost (Hewamalage et al., 2021).

Here, we assess the forecasting performance of XGBoost, LSTM, and traditional benchmark models for daily retail demand on a large, publicly available retail sales dataset. The models are evaluated in terms of their forecasting accuracy, feature importance, robustness to missing data and outliers, and computational efficiency. The results are intended to provide practical guidance for choosing suitable forecasting models to support inventory planning and supply chain decision-making in retail environments.

Literature Review

In this part, we shall look at how demand prediction methodologies in the retail sector evolved, from statistical analysis to machine learning to deep learning. We shall examine the drawbacks of traditional demand prediction methods, especially when dealing with non-linear data, and explore how gradient boosting algorithms, such as XGBoost, could improve on traditional models. We will also study deep learning methodologies, focusing on LSTM neural networks, and compare the data requirements of these models with those of tree-based ensembles.

Traditional and Machine Learning Methods

Demand forecasting has evolved significantly with the introduction of advanced analytical techniques (Azkaenza, 2023). Common methods such as exponential smoothing and ARIMA are widely used in retail because they are simple and easy to understand (Hyndman & Athanasopoulos, 2018). However, these methods are limited in their applicability, especially when data are nonlinear (Makridakis et al., 2020). Machine learning techniques have also been applied in various business contexts, particularly in nonlinear settings (Kumar & Minz, 2016). As Mayr et al. (2014) explain, boosting algorithms evolved from early machine learning classifiers (like AdaBoost) into statistical modeling tools that produce interpretable regression models. Gradient boosting is a popular machine learning method in business applications due to its strong performance with data-driven tasks (Natekin & Knoll, 2013). It builds a predictive model from a series of weak learners to reduce prediction errors (Natekin & Knoll, 2013). XGBoost, a gradient boosting algorithm, is widely used because of its ability to produce accurate models quickly (Chen & Guestrin, 2016). It uses a series of decision trees to minimize prediction error (Chen & Guestrin, 2016). Several studies have demonstrated the effectiveness of gradient boosting in retail, especially when incorporating calendar features and historical demand data (Huber & Stuckenschmidt, 2020). Tree-based ensemble methods are also valuable in retail because they provide feature importance, which supports decision-making (Huber & Stuckenschmidt, 2020).

Further empirical evidence has highlighted the efficiency of machine learning models, especially ensemble algorithms, in addressing complex retail demand dynamics influenced by exogenous factors such as price, promotion, and economic conditions. For example, Nasseri et al. (2023) evaluated various forecasting algorithms in the context of retail supply chains. They showed that ensemble tree algorithms, such as

XGBoost and gradient-boosted regression, performed exceptionally well across multiple datasets due to their ability to model nonlinear interactions between features. Similarly, Mitra et al. (2022) demonstrated that machine learning-based hybrid models that combine three algorithms with time-series models could improve forecasting precision in multichannel retailing. Therefore, machine learning models outperform conventional statistical models in most retail settings and can accommodate diverse data types.

Deep Learning and Comparative Performance (LSTM to Dataset Conditions)

Deep learning techniques have also proven effective for time-series forecasting tasks. Recurrent neural networks can learn sequential information using their hidden states. However, traditional recurrent networks struggle with vanishing gradients. To address the vanishing-gradient problem, Hochreiter and Schmidhuber (1997) introduced long short-term memory (LSTM) networks. These networks use memory cells to store long-term information, with their gates guiding this storage. These characteristics make long short-term memory networks more appropriate for solving sequential problems. However, studies suggest that deep learning techniques may require larger datasets to outperform traditional or machine learning methods. Hewamalage et al. (2021) suggested that deep learning techniques can outperform traditional or machine learning methods but may require larger datasets to achieve better performance. Studies on comparative forecasting techniques suggest that no model can outperform others across all conditions. Makridakis et al. (2020) indicated that a model's performance depends on the characteristics of the dataset used to train it. In retail forecasting, ensemble tree-based models can outperform deep learning techniques when the dataset is small or medium-sized. These traits of traditional or machine learning methods make them more suitable for comparative studies with deep learning techniques using real-world retail data.

Comparisons between deep learning and machine learning models suggest that their performance is influenced by data characteristics, feature extraction, and prediction horizons (Lei et al., 2024). The literature, including Theodoridis & Tsadiras (2024) and Niemelä (2025), shows that LSTM networks excel at recognizing temporal relationships in large, high-frequency datasets. However, XGBoost outperforms them on small retail datasets with features. Furthermore, studies have combined LSTM with boosting models to leverage the complementary nature of these methods, thereby improving robustness and precision (Joseph et al., 2022; Obi, 2024). These studies support the argument that there is no one-size-fits-all solution and that comparative analysis is important, especially in the retail sector, given differences in data sets.

Methodology

This section provides an overview of the methodology employed. Specifically, data description, data preparation, models, evaluation metrics, and robustness are discussed.

Data Description

The study employs the Store Item Demand Forecasting dataset from Kaggle. The dataset consists of 913,000 daily sales observations that were collected between January 1, 2013, and December 31, 2017. The data include individual-level sales of 10 retail stores and 50 products. Each observation is the sales quantity for a store-item combination on a given day. The dataset's large size and multi-year coverage make it appropriate for evaluating machine learning and deep learning models for retail demand forecasting.

Data Preparation

The data preprocessing was performed in Python using the pandas library. The Store Item Demand Forecasting dataset was first sorted chronologically by store, item, and date to preserve the temporal structure of the forecasting problem. The date variable was used to extract calendar variables, including

month, day of month, day of week, week of year, quarter, and a binary variable indicating weekend, to capture seasonal patterns in retail demand. In addition, lag features capturing sales from the past 1, 7, 14, and 28 days, and rolling mean features calculated over 7-, 14-, and 28-day windows, were created to account for short-term demand trends and temporal dependencies. Any rows with missing values resulting from the lag and rolling calculations were dropped before model training. The dataset was split chronologically, with observations from 2013–2016 used for training and those from 2017 used for testing. The final feature set consisted of store and item identifiers, calendar-based variables, lag features, and rolling demand statistics, enabling the models to capture both temporal patterns and recent demand behavior. The input features of the LSTM model were normalized to the [0, 1] range using Min-Max normalization. However, normalization was not required for the XGBoost model. The LSTM model was trained on a 7-day sliding window. The input sequence consisted of seven days of demand observations, and the output was the demand for the next day.

3XGBoost Model

XGBoost was implemented through the XGBoost package. This is a type of gradient boosting that builds an ensemble of decision trees to minimize a regularized loss function (Chen & Guestrin, 2016). Bayesian hyperparameter tuning was performed using the Optuna package (Akiba et al., 2019), optimizing the objective function with Root Mean Squared Error (RMSE) calculated over time-wise folds (Hyndman & Koehler, 2006). Hyperparameters tuned included:

- Number of boosting rounds
- Maximum tree depth
- Learning rate
- Subsample ratio
- Column sampling ratio

A total of 100 different hyperparameter configurations were evaluated.

LSTM Model

The LSTM model was built using TensorFlow with the Keras API (Kode, Elleithy, & Almazaydeh, 2024). It features two LSTM layers and a dense output layer (Kode, Elleithy, &. The first LSTM layer has 36 units and returns the full sequence, while the second LSTM layer has 18 units and returns only the last hidden state. Dropout layers were added between the LSTM layers to prevent overfitting during training. The model was trained with the Adam optimizer and mean squared error as the loss function.

Evaluation Metrics

Model performance was evaluated using three widely adopted metrics for demand forecasting:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2)

MAE is a measure of the average magnitude of forecast error; RMSE is more concerned with the magnitude of forecast error, while R^2 indicates the amount of variation in observed demand that can be explained by the model.

Robustness Experiments

The robustness of the forecast models was tested in two experiments using the training set. For the first experiment, missing data were simulated by randomly dropping 5%, 10%, and 15% of the data points from the training set. Then, the missing values were filled using linear interpolation before training the models. For the outlier case, random data points from the training set were selected, and their values were replaced with extreme values 5 standard deviations above the mean. In both cases, the test set was kept unchanged across all experiments.

Results

This section presents the findings from the forecast performance evaluation of the benchmarks, XGBoost, and the LSTM model on the retail demand dataset. First, forecast accuracy will be measured using mean absolute error (MAE), root-mean-square error (RMSE), and R^2. Next, the feature importance analysis will be conducted using the XGBoost model.

Model Performance

Table 1 presents the forecasting performance of theNaïve, XGBoost, Seasonal Naïve,7-Day Moving Average, and LSTM models on the test dataset.

Table 1. Model Performance Comparison

Model	MAE	RMSE	R^2
Naïve	11.245	15.389	0.762
Seasonal Naïve	8.841	11.597	0.865
7-Day Moving Average	8.980	11.860	0.859
XGBoost	**6.089**	**7.911**	**0.937**
LSTM	9.160	10.880	-0.089

The XGBoost model showed the best forecasting performance among all tested models. The model achieved the minimum MAE (6.089) and RMSE (7.911) and, at the same time, the maximum R^2 (0.937), indicating that it explains about 93.7% of the variance in the demand data. The Seasonal Naïve and 7-Day Moving Average models had decent forecasting accuracy; however, they were always outperformed by the XGBoost model. The LSTM model had slightly lower MAE and RMSE than the Naïve baseline; however, it had a negative R^2 (-0.089). Therefore, the best performance in the forecasting task is shown by the XGBoost model.

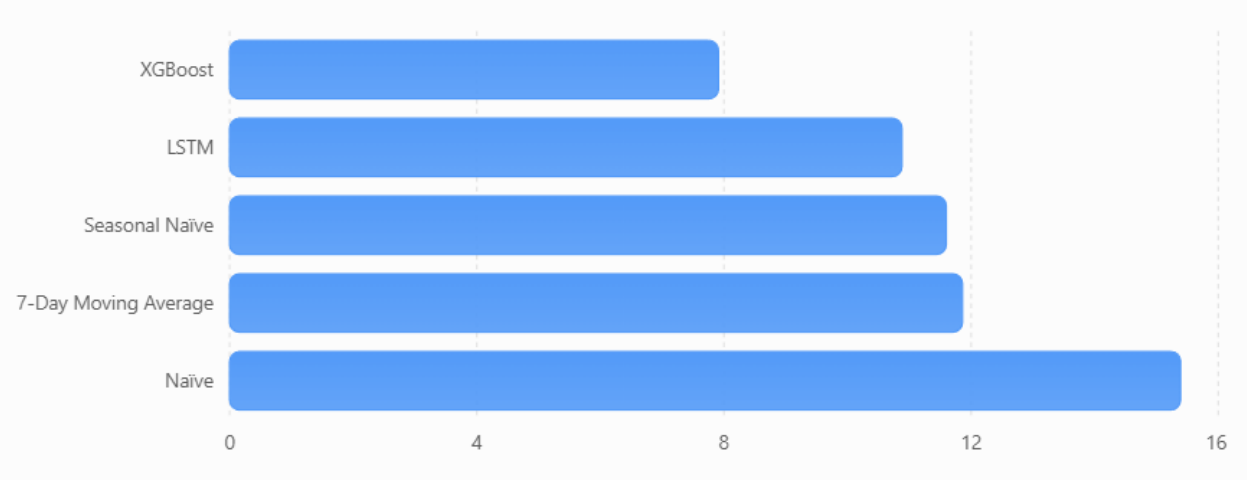


Figure 1: *RMSE comparison of the evaluated forecasting models.*

Figure 1 shows the RMSE values for all forecast models considered. The lowest RMSE value (7.911) was attained by the XGBoost model, which demonstrated the best forecasting precision. The Seasonal Naïve and 7-day Moving Average showed satisfactory results, while the Naïve forecast model exhibited the highest prediction error among all models. It should be noted that the LSTM model outperforms the Naïve model but remains less precise than the XGBoost model.

Feature Importance

Feature importance values derived from the XGBoost model are presented in Table 2.

Table 2: Feature Importance

Rank	Feature	Importance
1	rolling_7	**0.391**
2	lag_7	**0.250**
3	rolling_14	**0.226**
4	dayofweek	0.040
5	lag_14	0.029
6	lag_28	0.014
7	rolling_28	0.011
8	weekend	0.011
9	lag_1	0.010
10	month	0.010
11	weekofyear	0.004
12	day	0.004
13	store	0.0003
14	item	0.0003
15	quarter	0.0002

Table 2 shows the feature importance values obtained from the XGBoost model. The most influential predictor was the seven-day rolling average (0.391), followed by the seven-day lag (0.250) and the fourteen-day rolling average (0.226). The results suggest that the most important information in the recent demand history is for forecasting future sales. Calendar-related variables, such as day of the week, also improved prediction accuracy but were much less important. After adding lagged and rolling demand features, variables representing the store, item, and quarter had little effect.

Robustness Results

According to the robustness test, XGBoost remained stable and reliable in its forecasting ability, even in the presence of missing data or outliers. In fact, when 15 percent of the data was removed from the training dataset, the increase in RMSE was lower for XGBoost than for LSTM. The same happened with the outliers, as XGBoost proved more robust.

Computational Efficiency

Training the XGBoost model took about 30 minutes, while training the LSTM model took around 187 seconds. Such results demonstrate that gradient boosting is computationally efficient for retail demand forecasting.

Discussion

The experimental results indicate that the proposed XGBoost model outperformed the benchmark models and the Long Short-Term Memory model in forecast accuracy, interpretability, and computational efficiency. XGBoost had the lowest MAE (6.089) and RMSE (7.911) and explained approximately 93.7% of the variability in retail demand ($R^2 = 0.937$). While both the Seasonal Naïve and 7-Day Moving Average models achieved decent forecast accuracy, they consistently performed worse than the XGBoost model. Meanwhile, the proposed Long Short-Term Memory model yielded a negative value of R^2 (-0.089), which shows its poor generalization ability. Besides, XGBoost trained much faster than the Long Short-Term Memory model, thus making it a computationally efficient solution for forecasting.

The analysis of feature importance has shown that the most significant factors for future sales were demand-related attributes. Specifically, the 7-day rolling average (39.1%), 7-day lag (25.0%), and 14-day rolling average (22.6%) were the three most important features used in the model. Calendar attributes, such as the day of the week, also affected forecast performance; however, store, item, and quarter attributes had very little effect on forecast performance after accounting for temporal demand patterns. Thus, the results of this work show that the XGBoost model is a powerful tool for retail demand forecasting using historical sales data.

Summary of Findings

As the results show, the XGBoost model demonstrates superiority over the baseline forecasting models and the LSTM in terms of accuracy, robustness, and computational efficiency. This model shows the smallest forecasting errors and the greatest explanatory power, demonstrating its ability to capture demand patterns in the retail dataset. The feature importance analysis demonstrates that rolling demand statistics and lag features are the most important in forecasting future sales, while calendar features add additional information about seasonality.

Limitations and Future Research

There are several limitations of the study. First of all, despite the large size of the dataset – over 913 thousand daily sales observations, it is just one publicly available retail demand forecasting dataset that does not represent other retail sectors or business environments. Second, although external factors such as promotions, discounts, holidays, pricing policies, weather conditions, and stock levels may affect demand, they were not considered in the forecasting models. Third, there is only one deep learning architecture in

use – LSTM, but other neural network architectures might perform better. Fourth, the evaluation of computational efficiency was conducted on a single hardware setup, so that training time may differ in other computing environments.

Future research might include additional deep learning architectures, hybrid forecasting models, and explainable artificial intelligence techniques for improving model interpretability. Also, external demand drivers may be included, and the proposed approach may be validated using retail datasets from different industries and geographic locations.

Acknowledgment
Grammarly AI was used only for language clarity and proofreading. All academic content, analysis, and conclusions were developed independently by the authors.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the ACM SIGKDD Conference*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Azkaenza, M. (2023). *Employing machine learning techniques to forecast food inflation in Indonesia: The role of food price indices* (Master's thesis, Copenhagen Business School). https://research-api.cbs.dk/ws/portalfiles/portal/98733666/1584520_Employing_Machine_Learning_Techniques_to_Forecast_Food_Inflation_in_Indonesia_The_Role_of_Food_Price_Indices.pdf

Babai, M. Z., Ali, M. M., & Nikolopoulos, K. (2022). Impact of temporal aggregation on the stock control performance of intermittent-demand estimators. *Omega*, 106, 102524. https://doi.org/10.1016/j.omega.2011.09.004

Bergmeir, C., Hyndman, R. J., & Benítez, J. M. (2018). *Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation*. **International Journal of Forecasting, 34**(2), 303–312. https://doi.org/10.1016/j.ijforecast.2017.11.002

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD Conference*, 785–794. https://doi.org/10.1145/2939672.2939785

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1), 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. 10.1162/neco.1997.9.8.1735

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, 36(4), 1420–1438. ttps://doi.org/10.1016/j.ijforecast.2020.02.005

Hyndman, R. J., & Koehler, A. B. (2006). *Another look at measures of forecast accuracy*. International Journal of Forecasting**, 22**(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Joseph, R. V., Mohanty, A., Tyagi, S., & Mishra, S. (2022). A hybrid deep learning framework with CNN and Bi-directional LSTM for store item demand forecasting. *Computers & Industrial Engineering*. https://www.sciencedirect.com/science/article/pii/S0045790622005754

Kode, H., Elleithy, K., & Almazaydeh, L. (2024). Epileptic seizure detection in EEG signals using machine learning and deep learning techniques. *IEEE access*, *12*, 80657-80668.

Lei, C., Zhang, H., Wang, Z., & Miao, Q. (2024). Multi-model fusion demand forecasting framework based on attention mechanism. *Processes*, *12*(11), 2612.https://www.mdpi.com/2227-9717/12/11/2612/xml

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Mayr, A., Binder, H., Gefeller, O., & Schmid, M. (2014). The evolution of boosting algorithms: From machine learning to statistical modelling. *Methods of Information in Medicine*, 53(6), 419–427. DOI: 10.3414/ME13-01-0122: https://www.thieme-connect.com/products/ejournals/abstract/10.3414/ME13-01-0122

Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022). A comparative study of demand forecasting models for a multi-channel retail company: A novel hybrid machine learning approach. *Operations Research Forum*. https://doi.org/10.1007/s43069-022-00166-4

Nasseri, M., Falatouri, T., Brandtner, P., & Darbanian, F. (2023). Applying machine learning in retail demand prediction—A comparison of tree-based ensembles and long short-term memory-based deep learning. *Applied Sciences, 13*(19), 11112. https://www.mdpi.com/2076-3417/13/19/11112

Natekin, A., & Knoll, A. (2013). Gradient boosting machines: A tutorial. *Frontiers in Neurorobotics*, 7, 21. https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2013.00021/full?utm_source=pocket_shared

Niemelä, T. (2025). Demand forecasting in the retail environment: A comparative study of LightGBM, XGBoost, and MLP models. University of Vaasa. https://osuva.uwasa.fi/items/36a81ce5-6b36-4858-89a7-18d5eb05f695

Obi, C. (2024). Demand forecasting in retail business using the ensemble machine learning framework—A stacking approach. *American Scientific Research Journal*. https://www.researchgate.net/publication/384869869

Panhalkar, Y. U., Peshave, H., Malik, N., & Navghare, P. (2024). Demand forecasting using machine learning: A comparative study of LSTM, XGBoost, and temporal fusion transformer. https://www.researchgate.net/publication/400349399

Saha, S., Islam, M. K., Rahaman, M. A., Mondal, R. S., & Kamruzzaman, M. (2024). Machine learning driven analytics for national security operations: A wavelet–stochastic signal detection framework. *Journal of Computational Analysis and Applications*, *33*(8), 210.

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://link.springer.com/article/10.1007/s12351-020-00605-2

Theodoridis, G., & Tsadiras, A. (2024). Retail demand forecasting: A multivariate approach and comparison of boosting and deep learning methods. *International Journal on Artificial Intelligence Tools*. https://www.worldscientific.com/doi/abs/10.1142/S0218213024500015