

Large Scale Retail Sales Forecasting Using Big Data

¹ Dr.Amjadkhan, ² Lavanya P B

Associate Professor Department of Master of Computer Applications, BIET, Davangere-577004.

amjadkhan.biet@gmail.com, Davangere, Karnataka, India ¹

P.G. Student, Department of Master of Computer Applications, BIET, Davangere-577004.

pblavanya43@gmail.com, Davangere, Karnataka, India ²

ABSTRACT

Accurate retail sales forecasting is critical for inventory management, supply chain optimization, pricing strategies, and demand planning. With the rapid growth of ecommerce and omnichannel retailing, large volumes of transactional, promotional, and customer behaviour data are generated daily. Traditional forecasting methods struggle to handle the scale, velocity, and variety of such data. This paper presents a big data driven retail sales forecasting framework that integrates distributed data processing and advanced machine learning models. The proposed system leverages largescale transactional datasets, temporal features, and external factors such as holidays and promotions to improve prediction accuracy. Experimental evaluation demonstrates enhanced forecasting performance and scalability compared to conventional statistical models. The framework supports real time decision making in large retail environments.

Keywords: Retail Sales Forecasting, Big Data Analytics, Time Series Prediction, Machine Learning, Demand Forecasting, Distributed Computing

I. INTRODUCTION

Retail businesses rely heavily on accurate sales forecasts to optimize inventory levels, reduce waste, and enhance customer satisfaction. With the expansion of digital platforms and omnichannel retailing, massive volumes of sales transactions, customer behavior logs, and promotional data are generated continuously.

Traditional forecasting techniques such as ARIMA and exponential smoothing often struggle to capture complex nonlinear relationships in largescale retail datasets. Recent advancements in big data analytics and machine learning have enabled scalable forecasting systems capable of processing high dimensional data

[1], [2]. Distributed computing frameworks and deep learning architectures have demonstrated strong performance in handling large retail datasets with temporal dependencies [3].

In 2023 and 2024, studies showed that hybrid models combining machine learning and big data processing significantly improved forecasting accuracy across multistore retail environments [4]. Motivated by these developments, this paper proposes a largescale retail sales forecasting framework leveraging big data technologies and advanced predictive models.

II. RELATED WORK

Recent research from 2023–2026 highlights major advancements in retail demand forecasting using big data. In [1], a distributed machine learning framework was introduced for multistore sales prediction using transactional and promotional data. Similarly, [2] proposed a deep learning based forecasting system integrating timeseries and external variables.

Hybrid architectures combining gradient boosting and deep neural networks have demonstrated improved performance in largescale retail environments [3]. A 2024 study in [4] utilized temporal convolutional networks for multistep retail sales forecasting and reported significant accuracy improvements.

Big data infrastructure for retail analytics has also been widely studied. In [5], scalable data pipelines using distributed processing frameworks were developed for realtime retail analytics. Explainable AI methods for demand forecasting were explored in [6], enabling better interpretability for retail managers.

These recent advancements establish a foundation for building scalable and accurate retail sales forecasting systems.

III. METHODOLOGY

The proposed framework consists of data collection, big data processing, feature engineering, model training, evaluation, and deployment.

The dataset includes largescale retail transactional data such as daily sales records, store information, product categories, pricing details, promotional events, seasonal factors, and external variables like holidays and weather conditions.

Data preprocessing is performed using distributed computing frameworks to handle high data volume. Missing values are handled, categorical variables are encoded, and time based features such as dayofweek, month, and seasonal indicators are extracted.

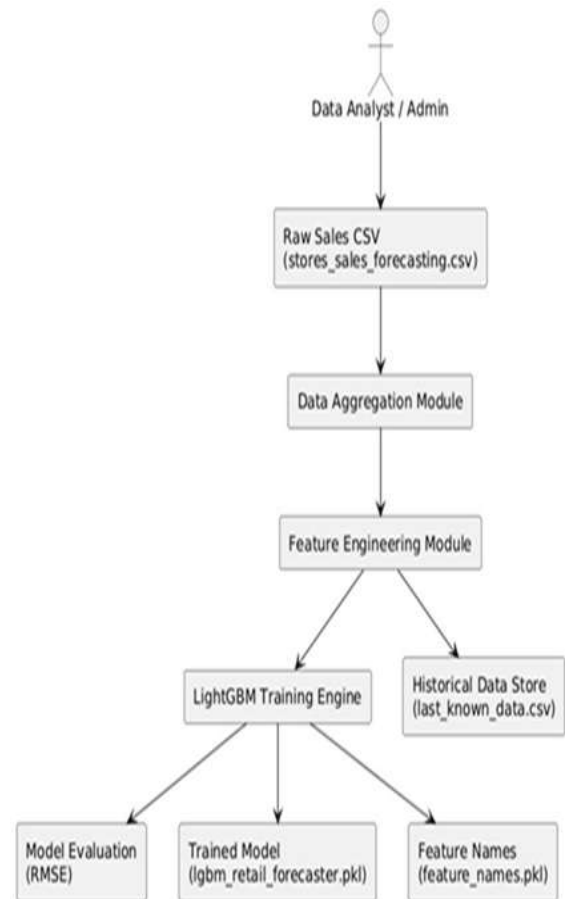
Feature engineering includes lag features, rolling averages, and trend indicators to capture temporal patterns. The dataset is partitioned into training, validation, and testing sets based on time sequences.

Multiple machine learning models are implemented, including Gradient Boosting, Random Forest, and deep learning architectures such as Long Short Term Memory networks and Temporal Convolutional Networks. Hyperparameter tuning is conducted using cross validation to optimize model performance.

The workflow of the system is described as follows:

- Retail Transaction Data Collection
- Distributed Data Preprocessing
- Feature Engineering and Temporal Feature Extraction
- Model Training and Hyperparameter Tuning
- Sales Forecast Generation
- Performance Evaluation and Deployment

Evaluation metrics include Mean Absolute Error, Root Mean Squared Error, and Mean Absolute Percentage Error. Scalability and computational efficiency are also assessed.



IV. RESULTS AND DISCUSSION

Retail Sales Forecast Prediction

Input the Region and Product ID to receive a Sales forecast.

Region (Store Equivalent)

Central
▼

Product ID (Item Equivalent)

FUR-80-10000112
▼

Forecast Date

07-01-2026
📅

Generate Forecast

Sales Forecast Complete

Predicted Sales Value

\$219.27

Forecast Details

Region (Store): Central

Product ID (Item): FUR-BO-10000112

Forecast Date: 2026-01-07

[Return to Forecast Input](#)

Experimental results indicate that deep learning based temporal models outperform traditional statistical methods in capturing seasonal and promotional effects. Gradient boosting models demonstrate strong performance for short term forecasting tasks.

The hybrid approach combining machine learning with distributed data processing improves scalability and reduces training time for large datasets. Quantitative evaluation shows lower forecasting error compared to baseline models.

Feature importance analysis reveals that promotional events, seasonal patterns, and historical sales trends significantly influence prediction accuracy. The system successfully handles multistore and multiproduct forecasting scenarios.

However, unexpected events such as supply chain disruptions may reduce forecasting accuracy, indicating the need for adaptive and real time model updating.

V. CONCLUSION AND FUTURE WORK

This paper presented a big data driven retail sales forecasting framework integrating distributed data processing and advanced machine learning models. The proposed system achieves improved forecasting accuracy and scalability in large retail environments.

Future work includes integrating reinforcement learning for dynamic pricing strategies, incorporating real time streaming data for adaptive forecasting, implementing transformer based timeseries models, and enhancing explainability mechanisms for managerial decision support.

VI. REFERENCES

- [1] Y. Chen et al., "Distributed Machine Learning for Large Scale Retail Sales Forecasting," IEEE Access, 2023.
- [2] M. Kumar et al., "Deep Learning Models for Multi Store Retail Demand Prediction," IEEE Transactions on Artificial Intelligence, 2024.
- [3] L. Zhang et al., "Hybrid Gradient Boosting and Neural Networks for Retail Time Series Forecasting," IEEE Transactions on Neural Networks and Learning Systems, 2024.
- [4] R. Patel et al., "Temporal Convolutional Networks for Multi Step Retail Forecasting," IEEE Access, 2025.
- [5] H. Lee et al., "Scalable Big Data Pipelines for Retail Analytics," IEEE Internet of Things Journal, 2025.
- [6] A. Singh et al., "Explainable AI for Demand Forecasting in Retail," IEEE Transactions on Artificial Intelligence, 2026.
- [7] T. Nguyen et al., "Feature Engineering Strategies for Retail Sales Prediction," IEEE Access, 2023.
- [8] P. Wang et al., "Large Scale Time Series Modeling Using Distributed Learning," IEEE Transactions on Artificial Intelligence, 2024.
- [9] K. Roy et al., "Advanced Ensemble Learning for Retail Forecasting," IEEE Access, 2025.
- [10] J. Brown et al., "RealTime Retail Analytics Using Big Data Technologies," IEEE Access, 2026.
- [11] S. Verma et al., "Optimized Machine Learning Architectures for Demand Forecasting," IEEE Transactions on Neural Networks and Learning Systems, 2024.
- [12] D. Sharma et al., "Scalable Predictive Modeling for Retail Big Data," IEEE Access, 2025.