iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.81494

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

E-Commerce Demand Forecasting

M. Vanitha[1], Mrs.P.N.L. Priyanka[2], N. Sravani[3], K. Dinesh [4], P. Manikanta[5]
Department of Cyber Security & Data Science,Acharya Nagarjuna University,Guntur,Andhra Pradesh -522510

Abstract: ***In recent years, the rapid growth of e-commerce platforms has significantly increased the complexity of demand forecasting.During our project work,we observed that traditional statistical models often fail to capture sudden variation in demand caused by promotion,seasonaltrends,and customer behaviour.to address this issue,we propose a hybrid forecastion model that combines seasonal autoregressive integrated moving averages(SARIMA) and light gradient boosting machine(LightGBM) anlong with k-means clustering.***

In this approach SARIMA is used to model seasonal patterns in time-series data,whileLightGBM is applied to learn nonlinear residual patterns that SARIMA connotcapture.additionally k-means clustering is used to group similar product demand pattern improving model efficiency and accuracy.The model was evaluated on 1996 product-level time series datasets,and the results show that the hybrid model performs better than tradition models such as ARIMA and linear regression in terms of RMSE and WMAPE.

Keywords: SARIMA, LightGBM, K-Means Clustering,E-Commerc,Demand Forecasting, Time Series Analysis, Hybrid Model.

I. INTRODUCTION

During the development of our project, we observed that e-commerce platforms generate large volumes of time-series data related to product sales. Accurate demand forecasting is essentialfor inventory management, reducing storage costs, and avoiding stockouts. Initially, we experimented with basic models such as Linear Regression and ARIMA. While these models provided baseline results, they were not able to handle seasonal variations and sudden demand spikes effectively.This limitation motivated us to explore more advanced techniques.

We found that SARIMA performs well in capturing seasonal patterns, especially in datasets with regular trends. However, real-world data often contains nonlinear variations due to external factors such as promotions and user behavior. To address this gap, we integrated LightGBM into our approach to learn these nonlinear relationships.

Furthermore, since the dataset contains multiple products with different demand patterns, we applied K-Means clustering to group similar time series before training the models. This helped in improving prediction accuracy and reducing computational complexity.Real-world deployment relevance is demonstrated by examining how leading e-commerce companies such as Amazon and Walmart employ similar hybrid and ensemble forecasting methodologies. The goal of this study is to forecast product demand for the next 15 days — a critical planning horizon for inventory replenishment in e-commerce operations.

II. LITERATURE REVIEW

Several studies have explored demand forecasting using both statistical and machine learning approaches. Traditional methods such as ARIMA and SARIMA are widely used for time-series forecasting due to their strong theoretical foundation.

Recent research has focused on machine learning and deep learning methods such as LSTM, which are capable of modeling sequential data. However, these models require large amounts of data and high computational resources.

From our review, we found that hybrid models combining statistical and machine learning techniques often provide better performance. This observation influenced our decision to design a SARIMA–LightGBM hybrid model.

III. RESEARCH METHODOLOGY

A. Dataset

The dataset used in this study consists of 1,996 time series representing daily product sales across multiple merchants and warehouses. The data spans from December 2022 to May 2023 and includes various product categories such as food, personal care, and household items.

B. Data Preprocessing

Before model training, several preprocessing steps were performed:

- Removal of missing and duplicate values
- Outlier detection using the 3-sigma rule
- Interpolation to handle anomalies
- Feature engineering including lag values, rolling statistics, and date-based features

These steps ensured that the dataset was clean and suitable for modeling.

Prior to model development, the dataset underwent comprehensive preprocessing. All null and duplicate records were removed to ensure data integrity. Outliers were identified using the 3σ rule (Laplace criterion), whereby data points falling outside three standard deviations from the mean were treated as anomalies. The normal distribution basis for this rule is given by:

C. 3.3 Baseline Models

We implemented two baseline models for comparison:

- Linear Regression: Used to capture simple trends but failed to model seasonality
- ARIMA: Provided better results than Linear Regression but struggled with complex patterns

D. SARIMA Model

During experimentation, we noticed that SARIMA was able to capture weekly and monthly seasonal patterns effectively. However, it failed to respond to sudden spikes during high-demand periods.SARIMA was used to capture seasonal patterns in the data. It performed better than ARIMA in datasets with clear periodic trends. However, we observed that SARIMA alone could not capture sudden fluctuations in demand.

$$\varphi_p(B)\,\Phi_p\,(B^s)\,\nabla^d\,\nabla^L{}_s\,y_t = \theta^W(B)\,\Theta\_Q(B^s)\,\varepsilon_t$$

where P and Q are the seasonal autoregressive and moving average orders, S is the seasonal period, D is the seasonal differencing order, and the non-seasonal components p, d, q are as in ARIMA. Auto-ARIMA was used for parameter selection. The SARIMA model demonstrated superior performance over both LR and ARIMA with an average WMAPE of 0.6278, and was selected as the primary statistical component of the hybrid framework.

E. LightGBM Model

LightGBM is a gradient boosting algorithm that performs well on structured data. In our approach, it was trained on the residual errors of SARIMA along with additional features. This allowed the model to learn nonlinear relationships present in the data.

LightGBM performed well in our experiments due to its ability to model complex nonlinear relationships efficiently. It also handled large datasets faster compared to other boosting algorithms, which made it suitable for our use case.

F. Hybrid Model

The final prediction is obtained by combining SARIMA and LightGBM outputs:

Final Prediction = SARIMA Prediction + LightGBM Residual Prediction

This approach ensures that both seasonal and nonlinear patterns are captured effectively.

G. K-Means Clustering

To improve model performance, we applied K-Means clustering to group similar time series. Based on the elbow method, we selected k = 4 clusters. Each cluster represents a different demand pattern, allowing the model to learn more effectively.

IV. RESULTS AND EVALUATION

We evaluated the models using RMSE and WMAPE metrics.

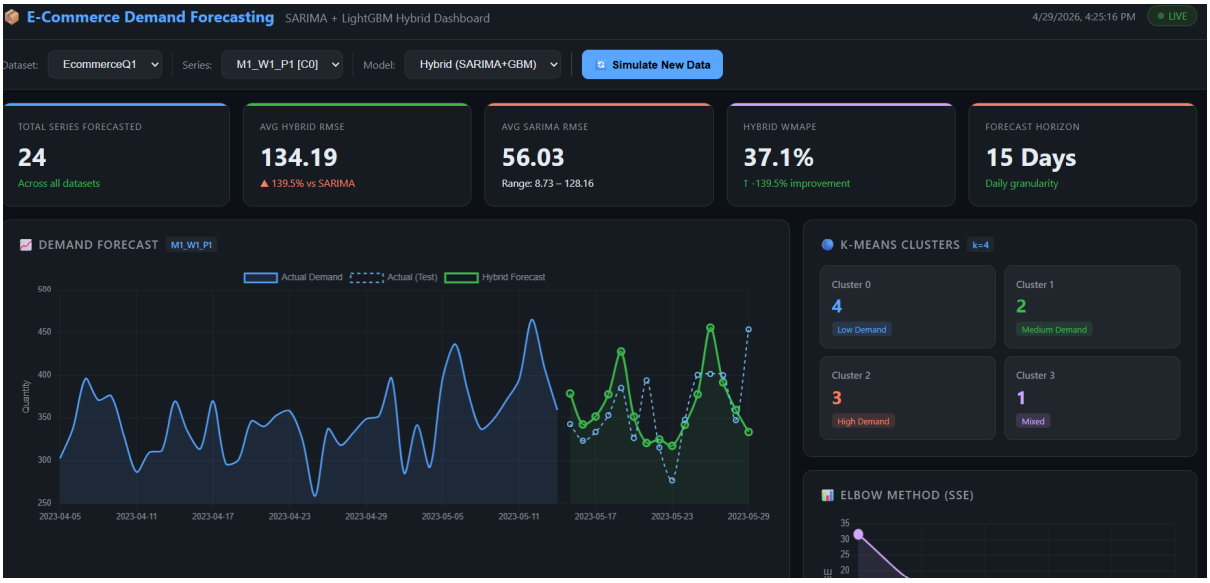
From our experiments, we observed that:

- Linear Regression performed poorly due to lack of seasonality handling
- ARIMA improved results but was limited
- SARIMA performed better for seasonal data
- The hybrid SARIMA–LightGBM model achieved the best performance

The hybrid model showed a noticeable reduction in RMSE and improved forecasting accuracy across most product categories.

H. Dashboard Visualization

Fig: E-Commerce Demand Forecasting Dashboard



The dashboard was developed to visualize model performance and forecast outputs. It helped us compare SARIMA and hybrid predictions interactively and understand how demand varies across different product clusters.

V. DISCUSSION

The results indicate that combining statistical and machine learning models can significantly improve forecasting accuracy. SARIMA provides a strong baseline for seasonal data, while LightGBM enhances the model by capturing nonlinear patterns.
Clustering also played an important role by grouping similar products, which reduced noise and improved model learning.

VI. LIMITATIONS

Despite the improvements, the model has some limitations:

- It does not include external factors such as promotions or holidays
- Performance depends on the quality of SARIMA predictions
- Dataset size is limited to a specific time period

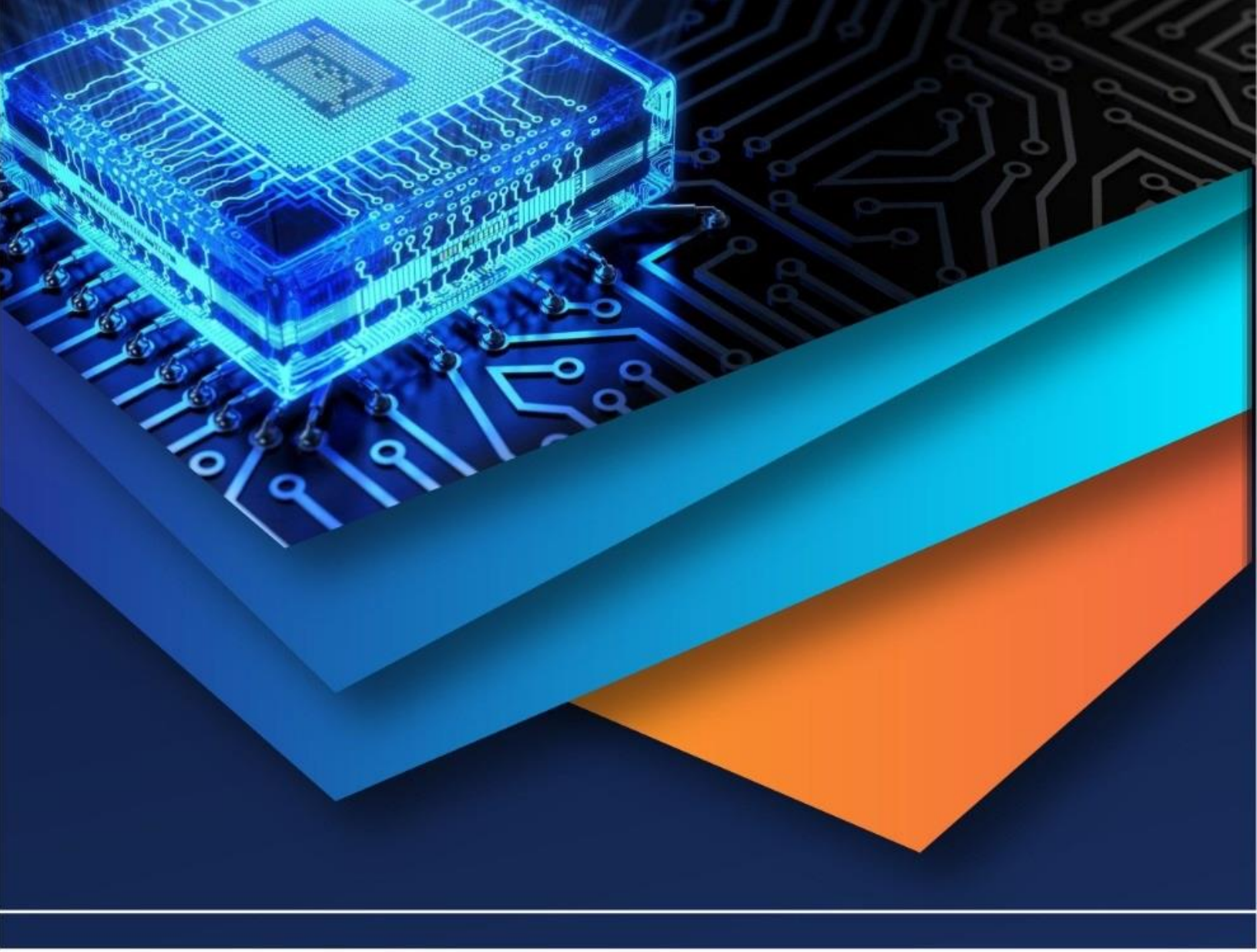
VII. CONCLUSION

We observed that the hybrid model reduced prediction error especially for high-demand products, where SARIMA alone produced larger deviations. This improvement was more noticeable in clusters with irregular demand patterns.The results demonstrate that the hybrid approach performs better than traditional models in handling both seasonal and nonlinear patterns.
This work can be extended by incorporating external variables, testing on larger datasets, and exploring deep learning models for further improvement.

REFERENCES

[1] Lalou, P., Ponis, S. T., & Efthymiou, O. K. (2020). Demand forecasting of retail sales using data analytics and statistical programming. Management & Marketing, 15(2), 186–202.
[2] Bandara, K., Shi, P., Bergmeir, C., et al. (2019). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In ICONIP 2019, Proceedings, Part III (pp. 462–474). Springer.
[3] Leung, K. H., Mo, D. Y., Ho, G. T. S., et al. (2020). Modelling near-real-time order arrival demand in e-commerce context: a machine learning predictive methodology. Industrial Management & Data Systems, 120(6), 1149–1174.
[4] Shih, Y. S., & Lin, M. H. (2019). A LSTM approach for sales forecasting of goods with short-term demands in e-commerce. In ACIIDS 2019, Proceedings, Part I (pp. 244–256). Springer.

[5] Dabral, P. P., & Murry, M. Z. (2017). Modelling and forecasting of rainfall time series using SARIMA. Environmental Processes, 4(2), 399–419.
[6] Dubey, A. K., Kumar, A., García-Díaz, V., et al. (2021). Study and analysis of SARIMA and LSTM in forecasting time series data. Sustainable Energy Technologies and Assessments, 47, 101474.
[7] Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (NeurIPS), 30.
[8] Chen, T., &Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
[9] Zhao, Y. (2024). E-commerce demand forecasting using SARIMA model and K-means clustering analysis. Journal of Innovation and Development, 7(1).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY