International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

ASales Forecasting & Demand Prediction System using ML Techniques

[1]A. Saipriya(student), [2]Mrs. G. Priyanka Assistant Professor(guide), [3]Mrs.P.Shraddha (Assistant Professor(HOD))

[1,2,3]KLR COLLEGE OF ENGINEERING AND TECHNOLOGY (Approved by AICTE ,New Delhi ,Affiliate to JNTU ,Hyderabad)

[1,2,3]B.C.M Road ,Paloncha ,BhadradriKothagudemDist.,Telangana,507115

[1]avulasaipriya123@gmail.com, [2]gadadesipriyanka@gmail.com, [3]shraddhaanair@gmail.com

***Abstract*— Accurate sales forecasting and demand prediction play a vital role in retail analytics by supporting inventory optimization, supply chain planning, and informed business decision-making. Conventional statistical forecasting techniques often struggle to capture complex, nonlinear sales patterns and dynamic market behavior, resulting in reduced prediction accuracy and inefficient resource utilization. A publicly available Amazon sales dataset containing historical transactional records with temporal, product, and category-related attributes was utilized to develop forecasting models for DailySales and DailyDemand. The workflow included dataset exploration, data cleaning, chronological sorting, exploratory data analysis, feature engineering, time-based train-test splitting, feature selection, standardization, model training, performance assessment, and explainable artificial intelligence visualization. Multiple regression algorithms, including Random Forest Regressor, XGBoost Regressor, MLP Regressor, Linear Regression, and a hybrid Voting Regressor, were implemented and comparatively analyzed. Model performance was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2 score, and Mean Absolute Percentage Error (MAPE). The Voting Regressor achieved superior forecasting accuracy, obtaining an R^2 score of 0.990 with a MAPE of 2.903 for DailySales prediction and an R^2 score of 0.995 with a MAPE of 2.985 for DailyDemand prediction. The integrated forecasting framework significantly enhances prediction reliability, enabling accurate category-wise sales and demand planning for intelligent retail management.**

"Keywords— Breast cancer diagnosis, Breast ultrasound imaging, YOLO object detection, Lesion segmentation, Morphological feature extraction, BI-RADS classification."

I. INTRODUCTION

Sales forecasting and demand prediction have become fundamental components of modern retail analytics, enabling organizations to optimize inventory levels, improve supply chain coordination, and enhance customer satisfaction. The rapid expansion of e-commerce platforms and digital retail ecosystems has generated large volumes of transactional data that provide valuable insights into consumer purchasing behavior and market trends. Accurate forecasting enables retailers to anticipate future demand, minimize operational costs, and allocate resources more efficiently while responding to changing customer preferences. With increasing product diversity and fluctuating market conditions, intelligent forecasting systems have emerged as essential decision-support tools for sustainable retail management and business growth [1]. The growing availability of historical sales data has further encouraged the adoption of data-driven forecasting approaches capable of supporting strategic planning and operational efficiency across multiple retail domains [2].

Despite considerable progress in forecasting methodologies, several challenges continue to limit the effectiveness of existing solutions. Many conventional forecasting techniques rely on simplified assumptions and often fail to represent complex relationships among sales patterns, seasonal variations, product categories, promotional activities, and changing consumer behavior. Furthermore, retail environments generate highly dynamic and heterogeneous data, making it difficult for traditional approaches to maintain consistent predictive performance over extended periods. Limited adaptability, reduced scalability, and insufficient support for category-level demand estimation often result in inaccurate forecasts, leading to stock shortages, excess inventory, and increased operational expenses [3]. These limitations emphasize the need for more reliable and intelligent forecasting frameworks capable of handling diverse retail scenarios while improving prediction consistency and business decision-making [4].

The primary objective is to develop an intelligent sales forecasting and demand prediction framework that supports accurate estimation of future retail performance using historical transactional information. The proposed framework aims to provide reliable forecasts for multiple product categories while enabling effective decision support for inventory planning and resource allocation. In addition, the framework is intended to offer an accessible deployment environment that allows users to generate forecasting results through a practical interface suitable for real-world retail applications. The overall contribution lies in integrating comprehensive forecasting capabilities with user-oriented decision support, thereby facilitating efficient planning and

improved operational management across different retail environments [5]. The developed framework also promotes scalable and flexible forecasting for diverse business requirements while maintaining prediction reliability and practical usability [6].

The significance of this contribution extends beyond forecasting accuracy by supporting informed business decisions that directly influence inventory optimization, procurement planning, and customer service quality. Reliable demand estimation enables organizations to reduce product shortages, minimize overstock situations, and improve overall supply chain efficiency, thereby enhancing profitability and resource utilization. Furthermore, the availability of accurate category-wise forecasts provides valuable insights that assist retailers in responding proactively to evolving market conditions and consumer preferences. Such intelligent forecasting capabilities contribute to more resilient retail operations while encouraging data-driven management practices across competitive business environments [7]. Consequently, the proposed framework represents a practical advancement toward efficient retail planning, sustainable inventory management, and enhanced organizational performance in modern commerce [8][9][10].

II. Related Work

Recent advances in retail analytics have significantly improved sales forecasting and demand prediction through the application of intelligent data-driven techniques. Khan et al. presented a business intelligence framework integrated with machine learning to enhance demand forecasting accuracy and support strategic retail decision-making, demonstrating the value of combining analytical platforms with predictive models for operational planning [11]. Punia and Shankar developed a deep learning-based decision support framework capable of modeling complex demand patterns and improving forecasting performance under dynamic market conditions, highlighting the potential of advanced predictive systems for retail applications [12]. Cadavid et al. conducted a comprehensive review of machine learning applications in demand and sales forecasting, emphasizing the increasing adoption of intelligent forecasting methods across industries while identifying challenges related to model selection, data quality, and practical deployment [13].

Subsequent investigations focused on improving forecasting accuracy and evaluating different predictive paradigms. Cadavid et al. further emphasized the importance of integrating domain knowledge with machine learning techniques to achieve robust forecasting performance across diverse industrial scenarios, while noting the limited availability of standardized evaluation practices [14]. Nasseri et al. compared tree-based ensemble methods with deep learning architectures for retail demand prediction and reported that model performance varied depending on data characteristics, indicating that no single forecasting approach consistently outperformed others under all conditions [15]. Cheriyan et al. demonstrated the applicability of machine learning techniques for intelligent sales prediction, showing promising improvements over conventional forecasting methods but acknowledging challenges associated with changing consumer behavior and evolving market dynamics [16].

Recent contributions have increasingly emphasized practical implementation and comparative evaluation. Panarese et al. introduced a machine learning-based sales forecasting platform designed to support industrial decision-making through an efficient predictive environment, illustrating the importance of integrating forecasting capabilities into business operations while recognizing the need for improved interpretability and scalability [17]. Jewel et al. performed a comparative analysis of multiple machine learning models for retail sales demand forecasting and demonstrated that forecasting performance differs considerably across datasets and evaluation scenarios, reinforcing the importance of selecting appropriate predictive frameworks for specific retail contexts [18]. Syberg et al. proposed a predictive sales and demand planning framework that incorporated enriched data for customer-oriented manufacturing environments, highlighting the benefits of combining diverse information sources while identifying challenges in adapting forecasting systems to changing operational requirements [19].

Although these contributions have substantially advanced intelligent forecasting, several limitations remain. Many existing solutions primarily emphasize predictive accuracy while providing limited support for integrated sales and demand estimation, user-oriented deployment, category-level forecasting, and practical decision support for inventory management. Differences in data characteristics, evaluation strategies, and implementation environments also make it difficult to generalize forecasting performance across retail applications. Malviya and Bhandari presented a systematic overview of machine learning approaches for demand prediction and concluded that future forecasting systems should improve reliability, adaptability, and real-world applicability while supporting comprehensive business decision-making [20]. Motivated by these observations, the present contribution develops an integrated forecasting framework that delivers accurate category-wise sales and demand estimation through a practical deployment environment, thereby enhancing operational planning, inventory optimization, and retail decision support.

III. Materials And Methods

The proposed system presents an intelligent machine learning framework for category-wise sales forecasting and demand prediction using historical Amazon sales data. A unified data processing pipeline is employed to prepare historical transactional records for predictive modeling, while advanced time-series feature engineering captures temporal dependencies through lag variables, moving averages, exponential moving averages, rolling statistics, rate-of-change indicators, cyclical representations, and category encoding. Multiple regression models, including Random Forest Regressor, XGBoost Regressor, MLP Regressor, Linear Regression, and an ensemble Voting Regressor, are developed and comparatively evaluated to identify the most reliable forecasting model. The ensemble approach enhances prediction robustness and generalization

by combining the strengths of multiple learners. Explainable artificial intelligence techniques, including SHAP and LIME, improve model transparency by identifying the features that most strongly influence forecasting outcomes. The selected models are integrated into a Flask-based web application that supports user authentication, CSV/Excel file uploads, and recursive seven-day forecasting, providing an efficient, interpretable, and practical decision-support platform for inventory planning and retail demand management.

Fig. 1. System Architecture

This end-to-end predictive pipeline processes an Amazon Sales Dataset through distinct data exploration, visualization, preprocessing, and feature engineering stages. The curated data undergoes an 80/20 train-test split to build and train multiple regression models, including Random Forest and XGBoost. Post-training, models are evaluated across key performance metrics, saved along with their scalers, and interpreted using SHAP and LIME for explainability. Finally, the optimized models are deployed via a Flask web application, allowing users to upload data and view interactive daily or category-wise forecasts.

A) Dataset Collection

The Amazon Sales Dataset, obtained from the Kaggle repository, is utilized to develop and evaluate the proposed sales forecasting and demand prediction framework. The dataset consists of historical daily retail transaction records containing temporal attributes, product category information, sales values, demand values, pricing details, and discount-related features. The forecasting targets are DailySales and DailyDemand, while additional time-series features are derived during preprocessing to capture temporal trends and category-specific behavior. The dataset covers multiple product categories with diverse sales patterns, making it suitable for training and evaluating machine learning models for accurate category-wise retail forecasting.

b) Pre-Processing:

The preprocessing stage prepares the sales dataset for accurate forecasting by improving data quality, generating meaningful features, creating prediction targets, and organizing the data to support reliable machine learning model development.

Data Pre-processing: Data pre-processing is performed to improve the quality, consistency, and chronological integrity of the sales dataset before model development. Initially, the order date is converted into a standardized date format, duplicate records are removed to eliminate redundant information, and the dataset is arranged in chronological order based on order date and product category. These operations ensure accurate temporal sequencing, reduce data inconsistencies, and establish a reliable foundation for time-series forecasting and demand prediction models.

Exploratory Data Analysis (EDA): Exploratory Data Analysis is conducted to examine the statistical characteristics and underlying patterns within the dataset before model training. Various analyses, including average price distribution, sales trends, demand trends, category-wise comparisons, and correlation analysis, are performed to understand relationships among variables and identify significant influencing factors. This preliminary investigation provides valuable insights into data behavior, supports informed feature selection, and facilitates the development of more accurate and reliable forecasting models.

Feature Engineering: Feature engineering is performed to transform the original dataset into a more informative representation for predictive modeling. Historical features, including lag values, moving averages, exponential moving averages, rolling standard deviation, and rate of change, are generated to capture temporal dependencies. Additionally, temporal, cyclical, signal-based, and categorical features are incorporated to represent seasonal variations and product characteristics. These engineered features enhance the learning capability of forecasting models, resulting in improved prediction accuracy and model generalization.

Target Variable Creation: Target variable creation is performed to prepare the dataset for supervised learning by defining future prediction objectives. Separate target variables are generated for sales forecasting and demand prediction by assigning the corresponding next-day values for each product category. This approach enables the forecasting models to learn temporal relationships between historical observations and future outcomes. Establishing well-defined target variables ensures consistent model training and improves the accuracy and reliability of future sales and demand predictions.

Data Normalization: Data normalization is applied to standardize the numerical feature values before model training, ensuring that variables with different scales contribute uniformly to the learning process. Normalization parameters are estimated exclusively from the training dataset and subsequently applied to the testing dataset to maintain evaluation integrity. Separate normalization models are maintained for sales and demand forecasting tasks, improving numerical stability, accelerating model convergence, and enhancing the predictive performance of machine learning algorithms.

C) Training and Testing:

A time-based train–test split is employed to preserve the chronological order of the dataset during model

development and evaluation. Earlier observations are allocated to the training dataset, while later observations are reserved for testing, ensuring that future information is not introduced during training. This strategy accurately reflects real-world forecasting scenarios, prevents information leakage, and enables objective assessment of model performance on previously unseen temporal data, thereby improving the credibility of the evaluation process.

D) Algorithms

Random Forest Regressor: Random Forest Regressor is an ensemble learning algorithm that generates multiple decision trees and combines their predictions to produce accurate regression results. Its ensemble strategy improves prediction stability, reduces overfitting, and effectively captures complex relationships among input features, resulting in robust and reliable forecasting performance.

XGBoost Regressor: XGBoost Regressor is a gradient boosting algorithm that sequentially constructs decision trees to minimize prediction errors. Its ability to model complex feature interactions, incorporate regularization, and optimize learning contributes to improved forecasting accuracy, enhanced generalization, and reliable predictive performance across diverse data patterns.

Linear Regression: Linear Regression is a supervised learning algorithm that estimates continuous target values by modeling linear relationships between input features and the response variable. Its computational efficiency, simplicity, and interpretability make it an effective baseline model for evaluating the performance of more advanced regression techniques.

MLP Regressor: MLP Regressor is a feed-forward artificial neural network that learns complex nonlinear relationships through multiple interconnected hidden layers. Its capability to model intricate feature interactions enhances prediction accuracy and enables effective learning of nonlinear patterns, improving the overall forecasting performance of regression tasks.

Extra Trees Regressor: Extra Trees Regressor is an ensemble algorithm that constructs multiple randomized decision trees to generate regression predictions. The incorporation of randomness increases model diversity, reduces prediction variance, and enhances robustness, making it an effective component for improving the stability and accuracy of ensemble forecasting models.

Voting Regressor: Voting Regressor is an ensemble technique that combines predictions from multiple regression models to generate a unified forecasting output. By aggregating the strengths of individual learners, it reduces prediction variability, improves robustness, and achieves higher forecasting accuracy than individual regression models operating independently.

Recursive Forecasting: Recursive forecasting is a multi-step prediction strategy that generates future forecasts sequentially by using previously predicted values as inputs for subsequent predictions. This iterative approach enables continuous long-term forecasting while maintaining temporal consistency, making it suitable for predicting future sales and demand over multiple time horizons.

StandardScaler: StandardScaler is a feature normalization technique that standardizes numerical variables by transforming them to a common scale. This preprocessing improves numerical stability, ensures balanced feature contributions, accelerates model convergence, and enhances the predictive performance of machine learning algorithms sensitive to feature magnitude.

IV. Experimental Results

MSE: Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (1)$$

RMSE: The root mean square error (RMSE) measures the average difference between a statistical model's predicted values and the actual values. Mathematically, it is the standard deviation of the residuals. Residuals represent the distance between the regression line and the data points.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}||y(i) - \hat{y}(i)||^2}{N}} \quad (2)$$

MAE: Absolute Error is the amount of error in your measurements. It is the difference between the measured value and "true" value. For example, if a scale states 90 pounds but you know your true weight is 89 pounds, then the scale has an absolute error of 90 lbs – 89 lbs = 1 lbs.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

R2 Score: The sum squared regression is the sum of the residuals squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (4)$$

mAP: Mean Average Precision (MAP) is a ranking quality metric. It considers the number of relevant recommendations and their position in the list. MAP at K is calculated as an arithmetic mean of the Average Precision (AP) at K across all users or queries.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \quad (5)$$

Table. 1. Sales Forecasting Performance Evaluation

ML Model	MSE	RMSE	MAE	R2-Score	MAPE

XGBoost	622950250669.622	789271.975	446373.113	0.968	8.208
RandomForest	698246731606.479	835611.591	494423.912	0.964	8.962
LinearRegression	876546478942.152	936240.610	532162.009	0.955	11.303
MLPRegressor	434603485029.951	659244.632	371382.101	0.978	7.184
Voting Regressor	194502746890.157	441024.656	205490.841	0.990	2.903

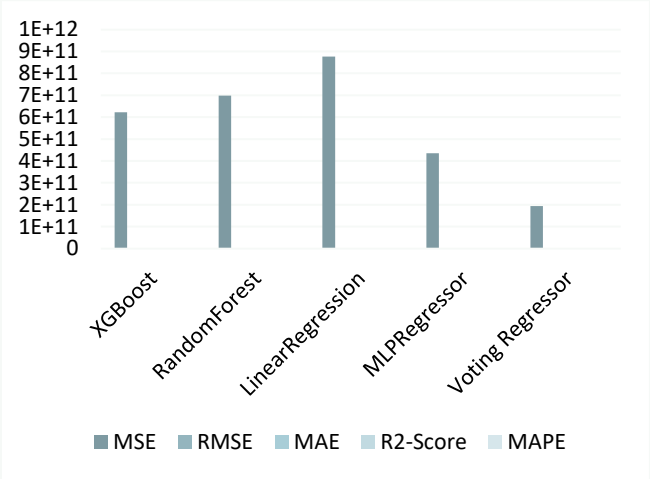
Table (1) presents the performance evaluation of regression models using MSE, RMSE, MAE, R2-Score, and MAPE, where Voting Regressor achieves superior accuracy and lowest error compared to other models considered.

Table. 1. Demand Prediction Performance Evaluation

ML Model	MSE	RMSE	MAE	R2-Score	MAPE
XGBoost	4340.078	65.879	40.939	0.980	9.170
RandomForest	8492.949	92.157	53.735	0.960	9.133
LinearRegression	7928.016	89.039	60.565	0.963	16.158
MLPRegressor	50214.817	224.087	93.636	0.765	7.703
Voting Regressor	967.352	31.102	15.824	0.995	2.985

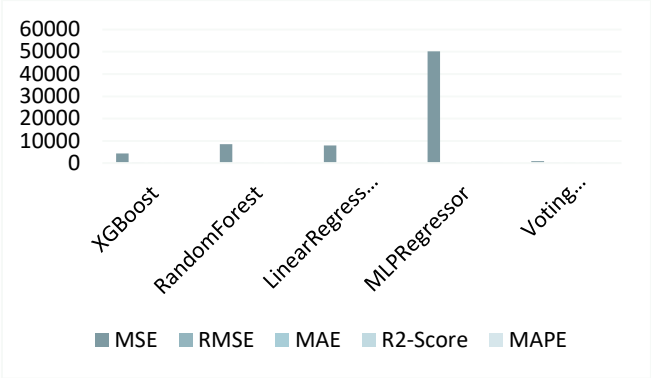
Table (2) presents performance evaluation of regression models using error metrics and R2-Score, where Voting Regressor achieves highest accuracy with lowest MSE, RMSE, MAE, and MAPE compared to others.

Graph. 1. Sales Forecasting Performance Evaluation



Graph. 1 evaluates five regression models across multiple performance metrics. Due to the massive y-axis scale, only the Mean Squared Error values are visibly distinct for each trained model.

Graph. 2. Demand Prediction Performance Evaluation



Graph. 2 evaluates five regression models across several performance metrics. The adjusted y-axis scale highlights a massive error spike for the MLP Regressor, whereas the Voting Regressor demonstrates the lowest overall error.

Fig.2. CSV/Excel Dataset Upload Page

Fig.2 showcases the web interface designed for dataset uploads. Users can import files like "samples.xlsx", utilize the "Configure Model" adjustments, and execute backend machine learning pipelines using the "Run Prediction" button.

Fig.3. Forecast Summary Dashboard

Fig.3 displays the forecasting results dashboard, presenting key predictive metrics and trend lines. "7-Day Cumulative Sales" shows the highest value at 56,677,735.7 Rupees, alongside interactive actual versus predicted charts for both sales and demand tracking.

V. CONCLUSION

In conclusion, the developed system successfully achieves the objective of providing accurate and reliable sales forecasting and demand prediction to support data-driven retail planning and inventory management. Using a historical Amazon sales dataset, the framework integrates comprehensive data preparation, time-series feature engineering, multiple regression models, comparative performance analysis, and explainable artificial intelligence to generate dependable forecasting outcomes for both DailySales and DailyDemand. Random Forest Regressor, XGBoost Regressor, Linear Regression, MLP Regressor, and an ensemble Voting Regressor were implemented and evaluated, with the Voting Regressor demonstrating the highest predictive performance. The selected model achieved an R^2 score of 0.995 for DailyDemand prediction, confirming the effectiveness and robustness of the proposed forecasting framework. To enhance practical usability, the trained forecasting models were deployed through a Flask-based web application that supports user authentication, CSV and Excel file uploads, category-wise one-day and seven-day recursive forecasting, and interactive visualization of prediction results. The integration of ensemble learning, explainable AI techniques, and web-based deployment provides a reliable and interpretable decision-support platform that enables automated forecasting, improves inventory planning, minimizes stock-related inefficiencies, and supports informed business decision-making in modern retail environments. The system can be further enhanced by incorporating real-time sales streams, external factors such as weather, holidays, economic indicators, and promotional campaigns to improve forecasting accuracy. Support for multi-store and multi-region forecasting can increase scalability for large retail enterprises. Integration with cloud-based platforms and automated inventory management systems can enable continuous forecasting and decision support. Additionally, incorporating advanced deep learning architectures and adaptive online learning mechanisms can improve model performance under rapidly changing market conditions.

REFERENCES

[1] Swami, S., Shah, A., and Ray, A. Predicting Future Sales of Retail Products using Machine Learning. arXiv:2008.07779.

[2] Haque, M., Amin, M., and Miah, S. Retail Demand Forecasting: A Comparative Study for Multivariate Time Series. arXiv:2308.11939.

[3] Salih, A. M., et al. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. arXiv:2305.02012.

[4] Douaioui, K., et al. Machine Learning and Deep Learning Models for Demand Forecasting in Supply Chain Management. ResearchGate, 2024.

[5] Hall, P. A Survey of Machine Learning Methods for Time Series Prediction. Applied Sciences, 2025.

[6] Theodoridis, T. A Comparative Analysis of Deep Neural Networks and the LSTMixer Architecture. Information, 2025.

[7] Scikit-learn Developers. Scikit-learn: Machine Learning in Python Documentation.

[8] Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System.

[9] Lundberg, S. M. and Lee, S. I. A Unified Approach to Interpreting Model Predictions.

[10] Ribeiro, M. T., Singh, S., and Guestrin, C. Why Should I Trust You? Explaining the Predictions of Any Classifier.

[11] Khan, M. A., Saqib, S., Alyas, T., Rehman, A. U., Saeed, Y., Zeb, A., ... & Mohamed, E. M. (2020). Effective demand forecasting model using business intelligence empowered with machine learning. IEEE access, 8, 116013-116023.

[12] Punia, S., & Shankar, S. (2022). Predictive analytics for demand forecasting: A deep learning-based decision support system. Knowledge-Based Systems, 258, 109956.

[13] Cadavid, J. P. U., Lamouri, S., & Grabot, B. (2018, July). Trends in machine learning applied to demand & sales forecasting: A review. In International conference on information systems, logistics and supply chain.

[14] Cadavid, J. P. U., Lamouri, S., & Grabot, B. (2018, July). Trends in machine learning applied to demand & sales forecasting: A review. In International conference on information systems, logistics and supply chain.

[15] 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.

[16] [Cheriyan, S., Ibrahim, S., Mohanan, S., & Treesa, S. (2018, August). Intelligent sales prediction using machine learning techniques. In 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE) (pp. 53-58). IEEE.

[17] Panarese, A., Settanni, G., Vitti, V., & Galiano, A. (2022). Developing and preliminary testing of a machine learning-based platform for sales forecasting using a gradient boosting approach. Applied Sciences, 12(21), 11054.

[18] [Jewel, R. M., Linkon, A. A., Shaima, M., Sarker, M. S. U., Shahid, R., Nabi, N., ... & Hossain, M. J. (2024). Comparative analysis of machine learning models for accurate retail sales demand forecasting. Journal of Computer Science and Technology Studies, 6(1), 204-210.

[19] Syberg, M., West, N., Lenze, D., & Deuse, J. (2023). Framework for predictive sales and demand planning in customer-oriented manufacturing systems using data enrichment and machine learning. Procedia CIRP, 120, 1107-1112.

[20] Malviya, P., & Bhandari, V. (2024). A systematic study on effective demand prediction using machine learning. Journal of Integrated Science and Technology, 12(1), 711-711.