

A COMPARATIVE MACHINE LEARNING FRAMEWORK FOR RETAIL SALES FORECASTING USING ENSEMBLE TECHNIQUE

Ramesh Kumar V
Department of CSE
C. Byregowda Institute of Technology (CBIT)
Kolar, Karnataka, India
rameshkv.eee@gmail.com

Manjunath Singh
Department of CSE
C. Byregowda Institute of Technology (CBIT)
Kolar, Karnataka, India

Anitha P
Department of CSE
C. Byregowda Institute of Technology (CBIT)
Kolar, Karnataka, India

Vanishree
Department of CSE
C. Byregowda Institute of Technology (CBIT)
Kolar, Karnataka, India

Abstract: Accurate forecasting of retail sales is essential for effective inventory management, supply chain coordination, and revenue optimization. However, modeling retail data is challenging due to inherent non-linearities and complex feature interactions. This study proposes a structured machine learning framework for product-level sales prediction using the BigMart dataset, integrating data preprocessing, feature engineering, and comparative model evaluation. Five regression models: Linear Regression, Support Vector Regression (SVR), Decision Tree, Random Forest and Extreme Gradient Boosting (XGBoost) are implemented and assessed using RMSE and R^2 metrics. Experimental results reveal that ensemble methods significantly outperform individual models, with XGBoost achieving the best performance ($R^2 \approx 0.924$) and the lowest prediction error. The study demonstrates that advanced ensemble techniques effectively capture intricate relationships among product attributes, pricing, and outlet characteristics. The proposed framework provides a scalable solution for retail analytics and supports the development of reliable decision-support systems

Keywords: Retail Sales Forecasting, Machine Learning, Ensemble Learning, XGBoost, Predictive Analytics

I. INTRODUCTION

In modern retail analytics, sales forecasting has become a key component of data-driven decision-making, as it directly affects inventory planning, supply chain operations, pricing policies, and overall business performance. Reliable

demand estimation allows organizations to optimize resource allocation and maintain efficiency across operations. However, forecasting accuracy is often challenged by the growing complexity of retail environments, influenced by factors such as evolving customer preferences, seasonal demand fluctuations, promotional campaigns, and competitive market dynamics. Traditional statistical methods are limited in this context, as they typically struggle to represent the non-linear and interactive relationships present in sales data [1], [2].

With the advancement of machine learning, more flexible and data-adaptive approaches have emerged for forecasting tasks. These techniques are capable of identifying complex patterns and dependencies within large datasets without relying on strict assumptions. Models such as Support Vector Regression (SVR), Decision Trees, Random Forests, and Extreme Gradient Boosting (XGBoost) have shown strong performance in capturing non-linear relationships and improving prediction accuracy in retail and e-commerce applications [3], [4]. Despite these developments, challenges remain due to data variability, the influence of external factors, and the limited availability of studies that provide a comprehensive comparison of multiple models, as many works focus on individual techniques [5].

II. LITERATURE REVIEW

Sales forecasting has attracted significant research attention due to its importance in retail analytics, particularly for inventory planning, demand estimation, and supply chain management. Traditional statistical approaches, including ARIMA and linear regression, have been widely applied in forecasting tasks. However, these methods often struggle to model non-linear patterns and complex interactions among variables, which are common in real-world retail datasets. As a result, machine learning-based techniques have increasingly been adopted to overcome these limitations and enhance predictive performance.

Recent studies have demonstrated the effectiveness of ensemble learning methods in improving forecasting accuracy. A recent machine learning framework (2025) reported that tree-based ensemble models, especially XGBoost, outperform conventional regression approaches by efficiently capturing feature interactions and reducing prediction errors [6]. Similarly, comparative analyses involving both machine learning and deep learning models have shown that ensemble and hybrid techniques achieve higher accuracy than standalone models, particularly when combined with effective feature engineering strategies [7].

Further research has highlighted the advantages of boosting algorithms in dynamic retail environments. A 2024 study found that XGBoost significantly reduces forecasting errors, such as RMSE and MAE, compared to traditional techniques, making it suitable for complex and high-variance datasets [8]. In addition, large-scale demand forecasting applications using datasets like Grupo Bimbo have shown that advanced models, including XGBoost and LightGBM, outperform classical approaches in handling high-dimensional and real-world supply chain data [9].

Comparative studies consistently confirm the superiority of ensemble-based approaches. Research conducted in 2025 evaluated multiple models, including Random Forest, Gradient Boosting, and time-series methods, and concluded that ensemble techniques deliver more reliable and accurate predictions across diverse retail scenarios [10]. Another study emphasized that preprocessing, feature engineering, and appropriate model selection play a crucial

role in improving forecasting performance, underscoring the need for a structured modeling pipeline [11].

Recent advancements have also introduced graph-based and deep learning approaches to capture dependencies across multiple retail locations, further improving prediction accuracy. Despite these developments, several challenges remain, including limited comparative evaluations across models, dataset dependency, and issues related to data variability and interpretability. These gaps highlight the necessity for a comprehensive framework that systematically evaluates multiple machine learning models under consistent conditions.

In response to these challenges, the present study proposes a comparative machine learning framework for retail sales forecasting using the BigMart dataset. The approach integrates data preprocessing, feature engineering, and model evaluation to analyze the performance of Linear Regression, Support Vector Regression, Decision Tree, Random Forest, and XGBoost. The objective is to identify the most effective model for capturing complex sales patterns and to provide a practical solution for real-world retail forecasting applications.

III. METHODOLOGY

This study presents a robust machine learning framework for retail sales forecasting that integrates data preprocessing, feature transformation, model development, hyperparameter optimization, and interpretability analysis. The BigMart dataset is utilized for training and evaluation. During preprocessing, missing values are addressed using statistical imputation techniques, with mean imputation applied to numerical features and mode imputation used for categorical variables to preserve data consistency. In addition, categorical inconsistencies are standardized to ensure uniform representation, and outliers are mitigated using the Interquartile Range (IQR) method to improve data quality and model stability. Feature engineering is employed to enhance predictive performance by deriving meaningful attributes from existing variables. In

particular, the variable `Outlet_Establishment_Year` is transformed into `Outlet_Age`, as defined in Eq. (1), which provides a more interpretable representation of store maturity and its impact on sales:.

$$\text{Outlet_Age} = 2025 - \text{Outlet_Establishment_Year} \quad (1)$$

The proposed framework adopts a comparative modeling strategy by implementing multiple regression algorithms with varying levels of complexity. Linear Regression is used as a baseline model and represents the relationship between input variables and the target as expressed in Eq. (2). While computationally efficient and interpretable, it assumes linearity and is limited in capturing complex feature interactions

$$y = \beta_0 + \sum_{i=1}^n \beta_i x_i + \epsilon \quad (2)$$

Support Vector Regression (SVR) extends this capability by introducing kernel-based transformations, enabling the mapping of input features into higher-dimensional spaces. The model optimizes the objective function shown in Eq. (3), balancing model complexity and error tolerance through margin maximization:

$$\min \frac{1}{2} \|\omega\|^2 \text{ subject to } |y_i - (w \cdot x_i + b)| \leq \epsilon \quad (3)$$

The Decision Tree Regressor partitions the feature space recursively based on impurity reduction, typically measured using Mean Squared Error (MSE) as defined in Eq. (4). Although effective in modeling non-linear relationships and interactions, decision trees are prone to overfitting without proper regularization:

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

To overcome this limitation, the Random Forest Regressor constructs an ensemble of decision trees using bootstrap aggregation and feature randomness. The final prediction is obtained by averaging individual tree outputs, as shown in Eq. (5), which reduces variance and improves generalization:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x) \quad (5)$$

Extreme Gradient Boosting (XGBoost) further enhances predictive performance through a gradient boosting mechanism that sequentially minimizes residual errors. The model optimizes a regularized objective function, as defined in Eq. (6), combining a differentiable loss function with a penalty term to control model complexity:

$$\mathcal{L}(\Phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (6)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (7)$$

Model performance is quantitatively evaluated using Root Mean Square Error (RMSE) and the coefficient of determination (R^2), as defined in Eq. (8) and Eq. (9), respectively. RMSE measures the average magnitude of prediction errors, while R^2 reflects the proportion of variance explained by the model:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

To further enhance model performance, hyperparameter optimization is carried out using Grid Search with cross-validation, ensuring robust parameter selection and minimizing overfitting. Additionally, feature importance analysis is performed for ensemble models, particularly Random Forest and XGBoost, to identify the most influential predictors. This not only improves model interpretability but also provides actionable insights for retail decision-making.

IV. RESULTS AND DISCUSSION

The BigMart dataset comprises 8,523 training instances and 5,681 test instances, containing both numerical and categorical attributes. Initial data inspection identified missing values in features such as Item_Weight and Outlet_Size. These were addressed using statistical imputation techniques, where the mean was applied to numerical attributes and the mode to categorical variables.

A. Data Analysis and Preprocessing

Exploratory Data Analysis (EDA) was conducted to understand feature distributions, detect anomalies, and examine relationships among variables. The distribution plots of numerical features (Fig. 4.1) indicate that Item_Outlet_Sales follows a right-skewed distribution, reflecting the common retail pattern where a limited number of products generate a disproportionately high share of total sales. Such skewness suggests the presence of outliers and non-uniform data distribution, which can influence model performance

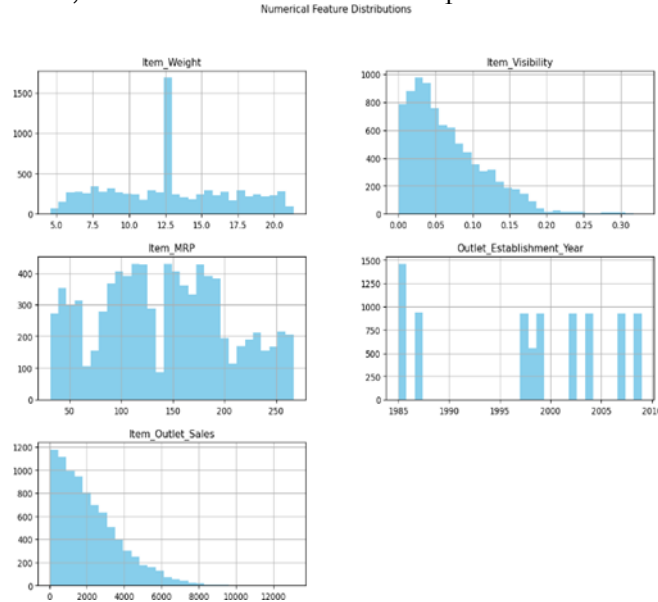


Fig. 4.1. Distribution plots of numerical features

Analysis of categorical variables (Fig. 4.2- Fig. 4.7) reveals variations in sales across different item categories and outlet types, indicating that these features contribute significantly to predictive modeling. The overall distribution of the target variable (Fig. 4.8) further reinforces the presence of skewness, highlighting the need for models capable of handling non-linear patterns and heteroscedasticity.

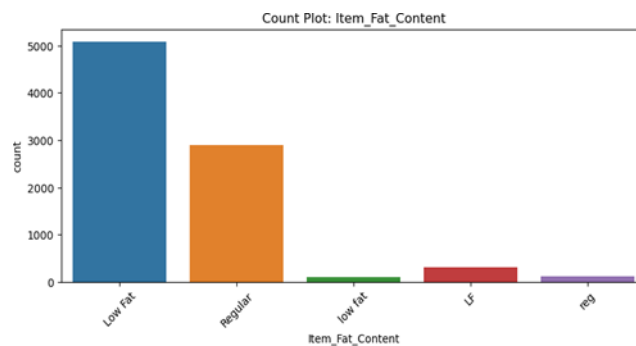


Fig.4.2. Count plot for feature Item_Fat_Content

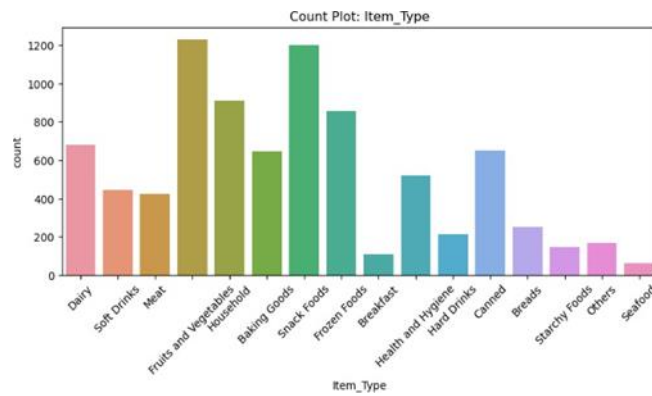


Fig.4.3. Count plot for feature Item_Type

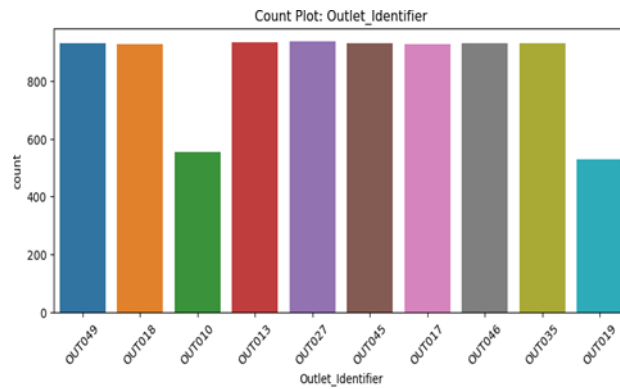


Fig.4.4. Count plot for feature Outlet_Identifier

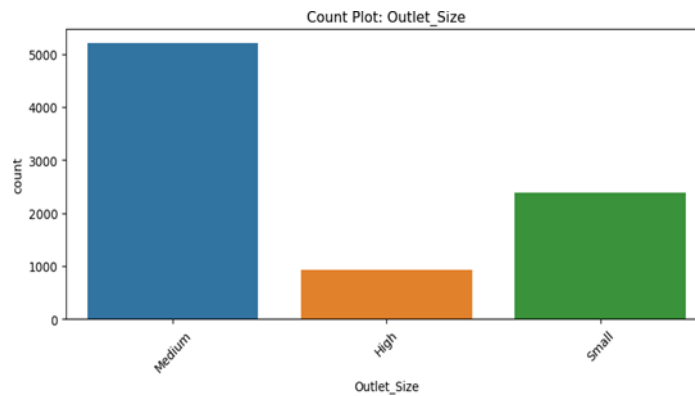


Fig.4.5. Count plot for feature Outlet_Size

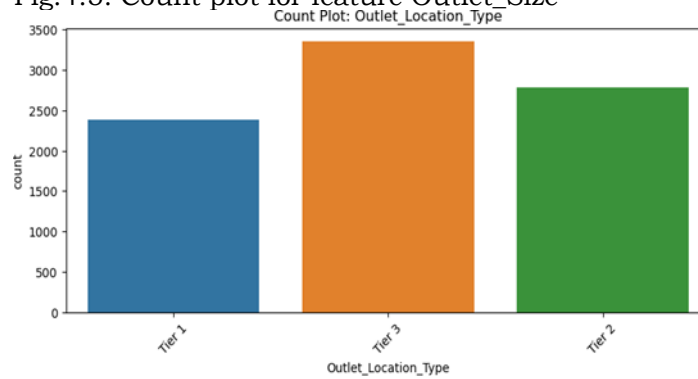


Fig.4.6. Count plot for feature Outlet_Location_Type

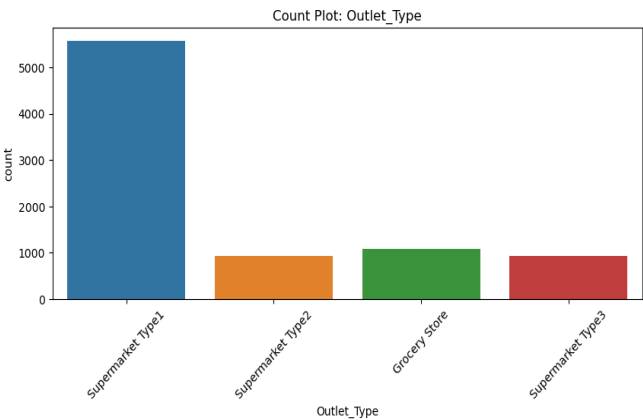


Fig.4.7. Count plot for feature Outlet_Type

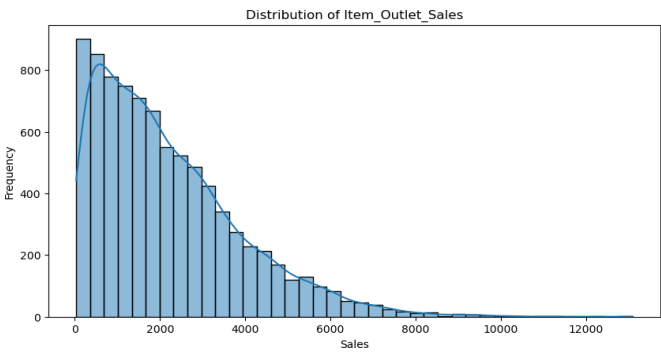


Fig. 4.8. Distribution plots target variable Item_Outlet_Sales

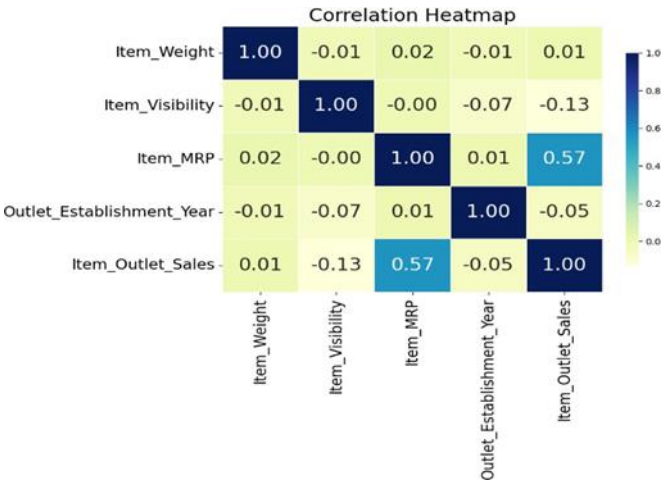


Fig. 4.9. Correlation matrix for numerical features

The correlation matrix (Fig. 4.9) provides a quantitative assessment of linear relationships between features. Among all variables, Item_MRP exhibits the strongest positive correlation with sales, suggesting that price is a key determinant of revenue. In contrast, most other features show weak to moderate correlations, indicating that linear models alone may not sufficiently capture the underlying data structure. To further investigate feature interactions, a pair plot

(Fig. 4.10) is analyzed, which visualizes pairwise relationships between numerical variables. The plot reveals scattered and non-linear patterns, along with varying data densities, confirming the presence of complex interactions among features. These observations justify the adoption of advanced machine learning models, particularly ensemble techniques, which are capable of capturing non-linear dependencies and improving predictive accuracy.

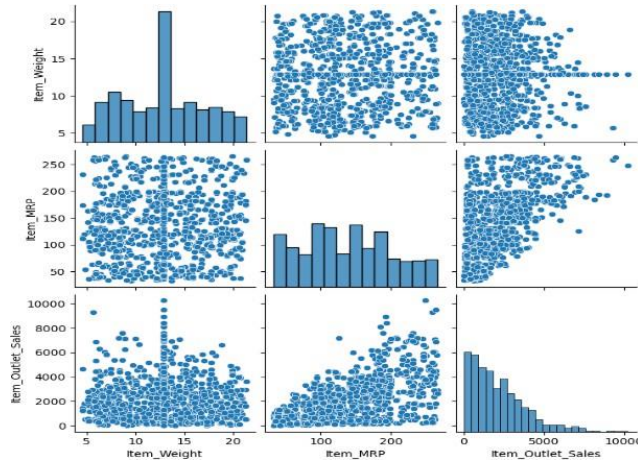


Fig. 4.10. Pair plot for numerical features

Overall, the preprocessing and exploratory analysis ensure data quality and provide critical insights into feature behavior, forming a strong foundation for subsequent model development and evaluation.

B. Model Performance and Discussion

The predictive performance of all implemented models is summarized in Table 1. A progressive improvement in model accuracy is observed with increasing model complexity and learning capability. Linear Regression, used as the baseline model, records the highest RMSE (1194.01) and the lowest R^2 value (0.510), indicating that it explains only about 51% of the variance in sales data. This performance limitation arises from its assumption of linear relationships, which is insufficient for modeling the complex and non-linear patterns present in retail datasets (Fig. 4.11).

Table 1. Comparison of various ML models

Model	RMSE	R^2
Linear Regression	1194.01	0.510
SVR	916.41	0.712
Decision Tree	912.90	0.714
Random Forest	551.74	0.895
XGBoost	470.61	0.924

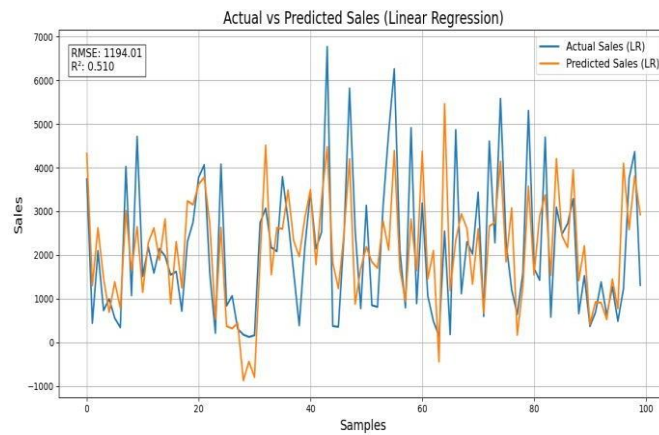


Fig. 4.11. Actual vs. Predicted Sales using the LR model.

Support Vector Regression (Fig. 4.12) and Decision Tree Regressor (Fig. 4.13) demonstrate moderate improvements, achieving RMSE values of 916.41 and 912.90, respectively, with corresponding R^2 scores above 0.71. These models are capable of capturing non-linear relationships; however, their effectiveness is constrained by certain limitations. SVR is sensitive to hyperparameter selection and kernel configuration, while Decision Trees are prone to overfitting, especially when model depth increases, which can negatively impact generalization performance.

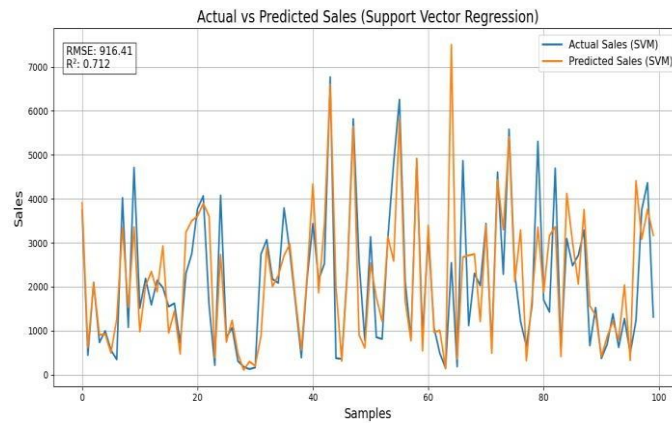


Fig. 4.12. Actual vs. Predicted Sales using the SVR model



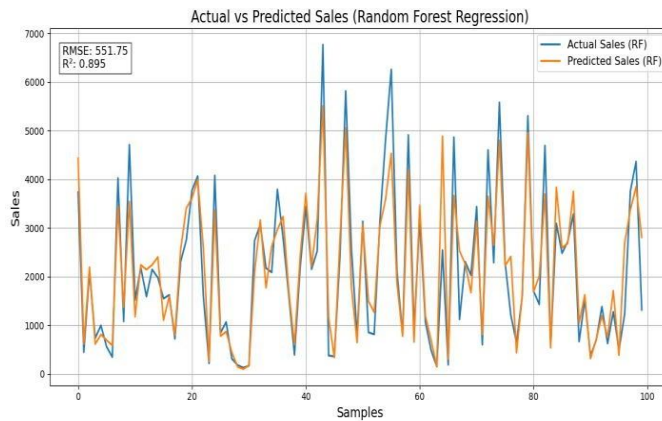


Fig. 4.13. Actual vs. Predicted Sales using the DTR model

Ensemble learning methods provide a significant enhancement in predictive accuracy. The Random Forest model (Fig. 4.14) achieves a substantial reduction in RMSE (551.74) and a high R^2 value (0.895), demonstrating its ability to reduce variance through aggregation of multiple decision trees. By incorporating bootstrap sampling and feature randomness, Random Forest improves robustness and stability compared to single-tree models

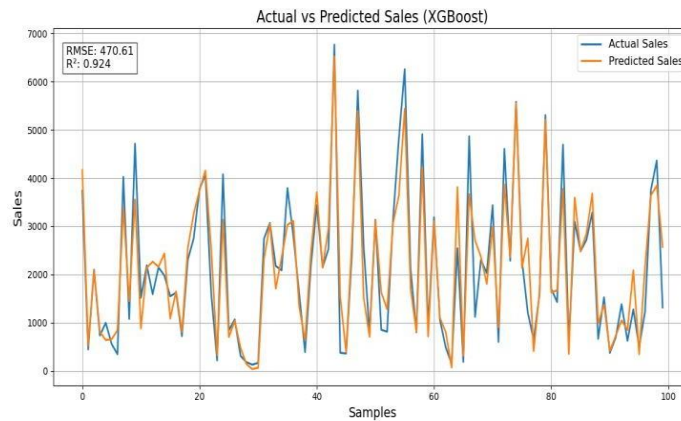


Fig. 4.14. Actual vs. Predicted Sales using the RFR model

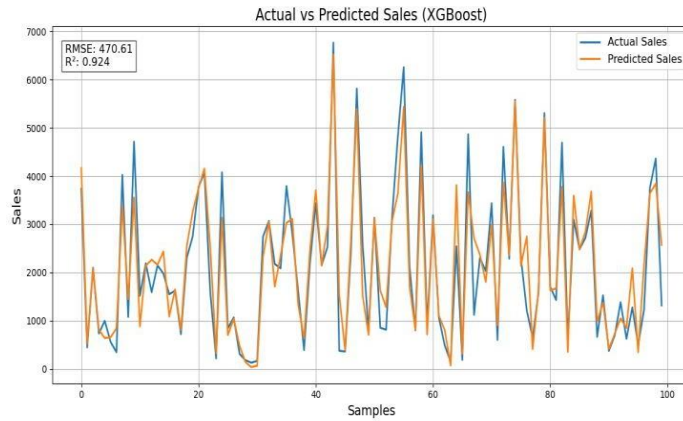


Fig. 4.15. Actual vs. Predicted Sales using the XGBoost model

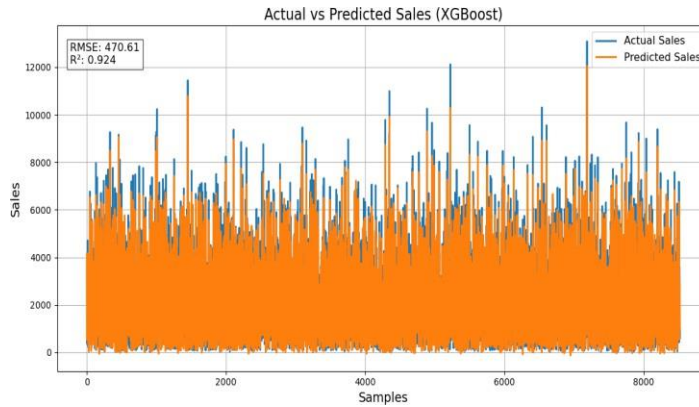


Fig. 4.16. Actual vs. Predicted Sales using the XGBoost model for full dataset

Among all models, XGBoost (Fig. 4.15 & Fig. 4.16) achieves the best performance, with the lowest RMSE (470.61) and the highest R^2 score (0.924), indicating that it explains over 92% of the variance in the target variable. This superior performance can be attributed to its gradient boosting mechanism, which iteratively minimizes residual errors, along with built-in regularization that controls model complexity and prevents overfitting. Additionally, XGBoost effectively captures complex feature interactions and non-linear dependencies, making it highly suitable for retail sales forecasting tasks.

VI. CONCLUSION

This study presented a machine learning framework for retail sales forecasting using the BigMart dataset, incorporating data preprocessing, feature engineering, and comparative model evaluation. Multiple regression models were analyzed using RMSE and R^2 metrics, revealing that traditional approaches are limited in handling non-linear patterns in sales data. In contrast, ensemble methods significantly improved predictive performance and generalization. Among

all models, XGBoost achieved the best results, demonstrating its effectiveness in capturing complex feature interactions and reducing prediction error. The proposed framework offers a reliable and scalable solution for retail forecasting and supports data-driven decision-making. Future work may explore deep learning models such as LSTM and Transformers for temporal modeling, inclusion of external factors like seasonality and promotions, and the use of Explainable AI techniques to improve interpretability. Additionally, extending the framework to real-time and hybrid systems can further enhance its practical applicability.

REFERENCES

- [1]. G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, Time Series Analysis: Forecasting and Control, 5th ed. Wiley, 2015..
- [2]. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," PLOS ONE, vol. 13, no. 3, 2018.
- [3]. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- [4]. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. ACM SIGKDD, 2016, pp. 785–794..
- [5]. A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. O'Reilly, 2019..
- [6]. Padmanabhan Venkateela, "Machine Learning Framework for Retail Sales Forecasting", International Journal of Computational and Experimental Science and Engineering, ISSN: 2149-9144, Vol. 11-No.4, pp. 7260-7269, 2025.
- [7]. Balusani, Akhil and Jai Ram Chowdary, Pathuri and Paruchuri, B. V. N. Prasad, Enhancing Retail Demand Forecasting with Xgboost: A Comparative Study with Machine Learning and Deep Learning Models. <http://dx.doi.org/10.2139/ssrn.5274510>
- [8]. Dankorpho, P. Sales Forecasting for Retail Business using XGBoost Algorithm. Journal of Computer Science and Technology Studies, 6(2), 136-141, 2024
- [9]. G. Ramkumar, M. Alluwaici, T. Vijetha, G. H, A. Venkatraman and R. M, "A Reliable System Performance Through NeuToMa with Spiking Neural Network Transformation in Network on Chip (NoC) Communication," 2026 International Conference on Communication, Computing and Emerging Technologies (IC3ET), Vasai, India, 2026, pp. 215-219, doi: 10.1109/IC3ET64989.2026.11467379.
- [10]. V. R, P. R. Kumar, D. Das, G. H, W. Hanbin and E. A, "Machine Learning-Driven Predictive Maintenance for Wind Turbines," 2026 International Conference on Electric Power and Renewable Energy (EPREC), Durg, India, 2026, pp. 1-6, doi: 10.1109/EPREC66546.2026.11412064.
- [11]. A. Tiwari, J. D. Kumar, W. Syed and G. H, "Efficient FPGA-Based CNN Accelerator for Real-Time Traffic Sign Classification on PYNQ-Z2," 2026 IEEE International Conference on AI Engineering and Innovations (AIEI), NIT Jamshedpur, India, 2026, pp. 1-4, doi: 10.1109/AIEI69164.2026.11496958.
- [12]. G. Ramkumar, M. Alluwaici, T. Vijetha, G. H, A. Venkatraman and R. M, "A Reliable System Performance Through NeuToMa with Spiking Neural Network Transformation in Network on Chip (NoC) Communication," 2026 International Conference on Communication, Computing and

- Emerging Technologies (IC3ET)*, Vasai, India, 2026, pp. 215-219, doi: 10.1109/IC3ET64989.2026.11467379.
- [13]. Girish H., Srinivas S., Kavya M., ...Kabbur S. "Swift Autonomous Fire Suppression Robot", 16th International Conference on Advances in Computing Control and Telecommunication Technologies Act 2025
 - [14]. Girish H., Soujanya S., "A Comprehensive Cadence-based Design Flow for RISC-V UART SOC IN SCL 180NM", 16th International Conference on Advances in Computing Control and Telecommunication Technologies Act 2025
 - [15]. Girish H., Rakesh S., Prasanna Raj C.P., Singh A.S., Yathisha L, "Deep Learning for Diabetic Retinopathy: Automated Detection and Classification from Retinal Images", 16th International Conference on Advances in Computing Control and Telecommunication Technologies Act 2025
 - [16]. Girish H., Shyam Sundar V., Priyanka C., Renuka Hebballi R.S, "RISC V Steel Processor Implementation for Communication between Two Chips using AXI Lite", 16th International Conference on Advances in Computing Control and Telecommunication Technologies Act 2025
 - [17]. M. H N, G. H and N. K N, "Simulation and Verification of SPI Master and Slave Communication," *2025 6th International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2025, pp. 1531-1534, doi: 10.1109/ICOSEC67334.2025.11459684. <https://ieeexplore.ieee.org/document/11459684>
 - [18]. S. S. V, P. C and G. H, "Design and Analysis of a D-Band LC Bandpass Filter and Low-Noise Amplifier for 6G Front-End Applications," *2026 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Gwalior, India, 2026, pp. 1-6, doi: 10.1109/IATMSI68868.2026.11466038. <https://ieeexplore.ieee.org/document/11466038>
 - [19]. "A Comprehensive Cadence-based Design Flow for RISC-V UART SOC IN SCL 180NM" , Girish H, Soujanya, Grenze International Journal of Engineering and Technology, 2025
 - [20]. "Swift Autonomous Fire Suppression Robot" Girish H et. Al., Grenze International Journal of Engineering and Technology, 2025
 - [21]. "Deep Learning for Diabetic Retinopathy: Automated Detection and Classification from Retinal Images" , Girish H, Cyril Prasanna Raj P , Grenze International Journal of Engineering and Technology, 2025
 - [22]. "RISC V Steel Processor Implementation for Communication between Two Chips using AXI Lite" Shyam Sundar V, Dr Girish H, Priyanka C, Renuka S Hebballi Grenze International Journal of Engineering and Technology, 2025
 - [23]. Sujatha.S, Mahendra BM, & Dr. Girish H. (2024). Smart Irrigation System for Optimal Water Conservation. *Revista Electronica De Veterinaria*, 25(1), 2613 - 2619. <https://doi.org/10.69980/redvet.v25i1.1311>
 - [24]. Mamatha K R, Thejaswini S, Rashmi N, & Dr. Girish H. (2024). Laser Fencing Surveillance System With Email Alert. *Educational Administration: Theory and Practice*, 30(2), 1539–1544. <https://doi.org/10.53555/kuey.v30i2.7919>
 - [25]. Thejaswini S, Mamatha K R, Rashmi N, & Dr. Girish H. (2024). Performance Evaluation of AI Models in Early Heart Disease Diagnosis. *Educational Administration: Theory and Practice*, 30(2), 1531–1538. <https://doi.org/10.53555/kuey.v30i2.7918>
 - [26]. Sushma K Sattigeri, Seema Srinivas, Shridhar Kabbur, Girish H, "Revolutionizing Urban Waste Management: An AI-Powered Autonomous

- Solution for Smart Cities”, Tuijin Jishu/Journal of Propulsion Technology ISSN: 1001-4055, Vol. 45 No. 02 (2024)
- [27]. Girish H, Vajjala Sai Sree Hari,Ponde Lakshmi Narasimha,P N Shreya,Dadigala Bhargavi, "THE DESIGN AND VERIFICATION OF A 32-BIT RISC V PROCESSOR USING VEDIC MATHEMATICS", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.12, Issue 5, pp.1130-1135, May 2024, Available at :<http://www.ijcrt.org/papers/IJCRTAB02172.pdf>
- [28]. IoT based home automation using Blynk cloud, J. Chetan Naik, B. S. Deeksha, H. Girish, Gurusharan Kashetty, T. G. Shivapanchakshari, Manohar Sai Polakam, Journal: AIP Conference Proceedings, AIP Conf. Proc. 2965, 020006 (2024), <https://doi.org/10.1063/5.0215427>, Published: July 2024
- [29]. IoT based fire alarm system using Blynk app, T. Bhargava, H. Kishor Kumar, Pramod Mahalingappa Godi, H. Girish, J. Chetan Naik, Journal: AIP Conference Proceedings, AIP Conf. Proc. 2965, 020007 (2024), <https://doi.org/10.1063/5.0212325>, Published: July 2024