National Conference on Information Technology & Digital Application (NCITSA-2025)
International Journal of Advanced Engineering Research and Science (IJAERS) Special Issue
https://dx.doi.org/10.22161/ijaers.ncitsa.2025.27 ISSN: 2349-6495(P) | 2456-1908(O)

Big Mart Sales Predictive Analysis using ML

Ms. Smita Agrawal[1] , Ishika Agarwal[2], Kanishk Khandelwal[3]

[1]Professor, Department of Information Technology, Jaipur Engineering College & Research Centre, Jaipur
[2]Student, Department of Information Technology, Jaipur Engineering College & Research Centre, Jaipur
Email: ishikaagarwal.it25@jecrc.ac.in
[3]Student, Department of Information Technology, Jaipur Engineering College & Research Centre, Jaipur
Email: kanishkkhandelwal.it25@jecrc.ac.in

Abstract*— *Accurate sales forecasting is essential for businesses to plan effectively, identify market trends, and improve profitability. Traditional methods, like tracking sales volume, are still used by retailers such as Big Mart but often lack precision. Incorporating Machine Learning (ML) techniques offers a more accurate and efficient solution. By mining customer and product data stored in data warehouses, ML models can uncover trends and predict future sales more reliably. Algorithms like Linear Regression, Ridge Regression, Polynomial Regression, and XGBoost outperform traditional models, enabling better inventory management and strategic planning. This shift helps businesses make informed, data-driven decisions and boost overall performance.

Keywords*— *Big Mart, Machine Learning, Predictive Analysis, XGBoost, Regression Models

I. INTRODUCTION

Big Mart, a prominent supermarket chain with stores across the country, has posed a challenge to data scientists to develop a model that can accurately predict product-wise sales for each store. The goal is to identify key products and store locations that significantly impact overall sales and leverage this insight to drive business success. With increasing competition from large retail chains and the rapid rise of global malls and online shopping platforms, the retail market has become highly competitive. To stay ahead, businesses aim to offer personalized and time-sensitive promotions, requiring accurate sales forecasts for effective stock control, logistics, and inventory planning.

Machine learning algorithms offer advanced capabilities for sales prediction, helping organizations address low-cost forecasting challenges. The dataset includes a combination of dependent and independent variables, covering product features, customer data, and inventory information. After preprocessing, this data is used to generate accurate predictions and uncover valuable insights. Techniques such as Random Forests and Linear Regression (both simple and multiple) are employed for forecasting. Evaluation metrics like Accuracy, MAE, MSE, and RMSE help determine the most effective algorithm for the task.

II. LITERATURE SURVEY

Numerous studies have been conducted to improve the accuracy of sales prediction in the retail domain using machine learning techniques, particularly for the Big Mart dataset.

Padma et al. [1] conducted a comparative study using Random Forest, Linear Regression, and Polynomial Regression. Among these, Random Forest achieved the highest accuracy due to its ability to handle non-linearity and prevent overfitting.

Hussain and Kalaimani [2] evaluated a variety of ML algorithms and reported that XGBoost outperformed other models in terms of accuracy. Their findings highlighted the importance of gradient boosting techniques in retail forecasting.

Gunjal et al. [3] proposed a framework that incorporated Generalized Linear Models (GLL), Gradient Boosted Trees (GBT), and Decision Trees. Their analysis concluded that GBT yielded the best prediction performance for Big Mart sales.

In an IEEE conference paper, Ranjitha and Spandana [4] explored XGBoost, Linear Regression, Polynomial Regression, and Ridge Regression. XGBoost demonstrated superior results, confirming its strength in handling complex retail datasets.

Behera and Nain [5], in their study published by Springer, focused specifically on XGBoost and evaluated its performance using various metrics. Their results reaffirmed the robustness and predictive power of the XGBoost algorithm for Big Mart sales data.

https://dx.doi.org/10.22161/ijaers.ncitsa.2025.27 ISSN: 2349-6495(P) | 2456-1908(O)

These studies collectively suggest that while traditional algorithms offer baseline performance, gradient boosting models—particularly XGBoost—consistently deliver higher accuracy in retail sales prediction tasks.

III. PROPOSED SYSTEM

The proposed system gives the most effective predictive analytics solution for sales forecasting, realizing the intended model's architecture illustration, which focuses on the colorful algorithm operations on the dataset. We calculate the delicacy, MAE, MSE, and RMSE in this stage before choosing the stylish yield algorithm. Our approach involves experimenting with different algorithms, including XGBoost, Linear regression, Polynomial regression, and Ridge regression, to identify the most effective technique for forecasting sales. Each algorithm offers unique strengths and capabilities, allowing us to explore different modeling approaches and select the one that best suits the characteristics of the sales data. During model training, we split the dataset into training and validation sets to evaluate the performance of each model and fine-tune its parameters for optimal results. Once the predictive models are trained, we evaluate their performance using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

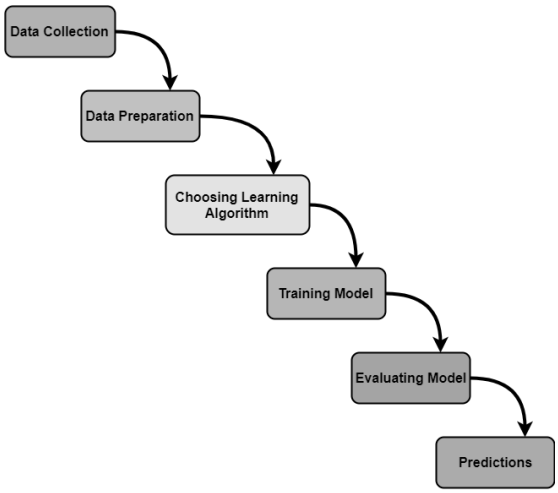

Fig. 1: Methodology

IV. OVERVIEW OF THE ALGORITHMS

To analyze and predict Big Mart sales data, we implemented and compared several machine learning algorithms, each with its strengths and limitations:

4.1 Linear Regression

Linear regression is a data analysis technique used to predict the value of an unknown variable based on a known, related variable. The first step involves building a scatter plot to observe the data pattern—whether linear or non-linear—and checking for variance or outliers. If the relationship isn't linear or outliers are present, data transformation may be necessary, though outliers should only be removed with valid, non-statistical justification. Next, the data is linked to the least squares regression line, and model assumptions are verified using a residual plot (to assess constant standard deviation) and a normal probability plot (to confirm the assumption of normality). If these assumptions are violated, transformation of the data might be required. Once the data is transformed, a new regression line is constructed using the least squares method. If a transformation was applied, the process should be revisited to ensure all conditions are met. If not, the process proceeds to the next phase. Once a "good fit" model is established, the final step is to write the least squares regression equation, including the standard estimation, predictions, and R-squared error values to evaluate the model's performance.

4.2 Polynomial Regression

Polynomial Regression is a regression technique that models the relationship between a dependent variable (y) and an independent variable (x) as nth-degree polynomial. The general equation for polynomial regression is:

https://dx.doi.org/10.22161/ijaers.ncitsa.2025.27 ISSN: 2349-6495(P) | 2456-1908(O)

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + \ldots + b_nx^n \quad (1)$$

where $b_0, b_1, b_2, \ldots, b_n$ are the coefficients. It is often considered a special case of multiple linear regression in machine learning because it involves adding polynomial terms to the linear regression equation to better capture the relationship and improve prediction accuracy. Polynomial regression is particularly useful when the dataset is non-linear in nature, as it allows the model to fit more complex patterns. Although it uses the linear regression framework, it can effectively capture the curves in the data by incorporating higher-degree terms, making it suitable for modeling non-linear relationships in a wide range of datasets.

4.3 Ridge Regression

Ridge Regression is a powerful model tuning technique used to handle datasets that suffer from multicollinearity, a condition where independent variables are highly correlated with each other. In such cases, traditional linear regression models can become unstable, producing large variances in the coefficient estimates. This instability leads to predictions that deviate significantly from the actual values, as the model becomes overly sensitive to minor changes in the data.

To address this issue, Ridge Regression applies L2 regularization, which involves adding a penalty term to the loss function equal to the square of the magnitude of the coefficients. This penalty term shrinks the coefficients towards zero, but unlike Lasso regression, it does not eliminate them entirely. By doing so, Ridge Regression reduces the model complexity and helps prevent overfitting, thereby improving the model's generalization performance. It ensures that even when multicollinearity exists, the estimates remain more reliable and closer to the actual values. Ridge Regression is especially useful in scenarios with a large number of features or when the features are not entirely independent of each other.

4.4 XGBoost Regression

Extreme Gradient Boosting, commonly known as XGBoost, is an advanced and highly efficient implementation of the traditional Gradient Boosting algorithm. While it follows the same fundamental principles, XGBoost significantly enhances performance by offering both a linear model solver and a tree-based learning algorithm. This dual approach makes it considerably faster—often several times quicker—than other existing gradient boosting methods. One of the key strengths of XGBoost is its support for a wide range of objective functions, including regression, classification, and ranking tasks, making it versatile across various machine learning problems. Due to its high predictive power, XGBoost is widely used in data science competitions and real-world applications where accuracy is critical.

Despite its complexity, XGBoost is designed for speed and performance, incorporating system optimization features such as parallel processing, cache awareness, and regularization to prevent overfitting. It also includes built-in functionality for cross-validation and feature importance analysis, which helps in model tuning and understanding the most influential variables in the dataset. This makes XGBoost not only powerful but also practical for large-scale machine learning tasks where both accuracy and computational efficiency are crucial.

V. CONCLUSION

This paper demonstrates that the XGBoost algorithm is superior for predictive analysis in retail sales, specifically in the context of Big Mart data. Its gradient boosting approach, regularization, and scalability contribute to its excellent performance. XGBoost not only reduces prediction error but also enhances interpretability through feature importance. This method can be adopted by retail chains for more accurate forecasting and better resource planning.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to our mentors and the Department of Information Technology at Jaipur Engineering College and Research Centre for their guidance and support throughout the preparation of this paper.

REFERENCES

https://dx.doi.org/10.22161/ijaers.ncitsa.2025.27 ISSN: 2349-6495(P) | 2456-1908(O)

[1] Padma, G., Chandana, S., Romitha, V., Reddy, Y. V., & Sukanya, M. (2023). Predictive analysis for Big Mart sales using machine learning algorithms. *Journal of Nonlinear Analysis and Optimization, 14*(2), 142–148.

[2] Hussain, S., & Kalaimani, G. (2022, October). Predictive analysis for Big Mart sales using ML algorithms. *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), 9*(5), 282–289.

[3] Gunjal, S. N., Kshirsagar, D. B., Dange, B. J., Khodke, H. E., & Kulkarni, C. S. (2022, May). Machine learning approach for Big-Mart sales prediction framework. *International Journal of Innovative Technology and Exploring Engineering, 11*(6), 69–75.

[4] Ranjitha, P., & Spandana, M. (2021, May). Predictive analysis for Big Mart sales using machine learning algorithms. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1416–1421). IEEE.

[5] Behera, G., & Nain, N. (2020, March). A comparative study of Big Mart sales prediction. In *Computer Vision and Image Processing (CVIP 2019), Communications in Computer and Information Science* (Vol. 1147, pp. 421–432). Springer.