IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141024

"Big Mart Sales Prediction Using Machine Learning"

Mr. Pavan Harilal Sonawane[1], Prof. Manoj Vasant Nikum*[2]

Research Scholar, Master of Computer Applications, Shri JRIT, Dondaicha, KBC NMU Jalgaon, Maharashtra, India[1]

Assistant Professor and HOD, Master of Computer Applications, Shri JRIT, Dondaicha, KBC NMU Jalgaon, Maharashtra, India*[2]

Abstract: Retail sales prediction plays a crucial role in effective inventory management, marketing strategy, and business profitability. This research focuses on predicting the sales of Big Mart outlets using various machine learning techniques. The dataset contains information on different products and store attributes. We compare algorithms such as Linear Regression, Random Forest, and XGBoost to determine the best-performing model for accurate sales forecasting. The results show that ensemble-based methods outperform traditional regression models in prediction accuracy.

Keywords: Sales Prediction, Big Mart, Machine Learning, Random Forest, Regression, Retail Analytics

1. INTRODUCTION

Sales forecasting helps retail businesses anticipate demand and optimize operations. Big Mart, a retail chain with multiple stores, faces challenges in predicting product-level sales due to varying store sizes, product types, and customer demographics.

Machine Learning (ML) provides data-driven solutions to these challenges by identifying hidden patterns and relationships within data. This paper explores different ML algorithms to predict item sales based on historical data and store-related factors.

2. LITERATURE SURVEY

1. Several studies have explored the use of machine learning in retail sales prediction:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141024

2. **Sharma et al. (2021)** used Linear Regression and Decision Trees to forecast sales, achieving moderate accuracy.

3. **Gupta & Singh (2022)** demonstrated that Random Forest yields better performance due to its ensemble nature.

4. **Kaggle BigMart Dataset (2023)** inspired multiple models combining data cleaning, feature engineering, and model tuning for improved accuracy.

5. This study builds upon these findings, focusing on model optimization and comparison.

3. RESEARCH METHODOLOGY

The main objective is to develop a predictive model that accurately estimates the sales of products in different outlets of Big Mart based on historical data and product/outlet attributes.

Specific goals:

- Identify key factors that affect sales.
- Handle missing, inconsistent, and categorical data effectively.
- Compare various machine learning models for prediction accuracy.
- Provide actionable insights for sales optimization.

3.1 Research Design

This study adopts a **quantitative and predictive research approach**, using **machine learning techniques** to forecast sales values based on historical data.
The design is **exploratory** (to understand relationships among variables) and **applied** (to solve a real-world business problem).

3.2 Data Collection Method

Data collection is a crucial step in building an accurate and reliable sales prediction model. In this study, **secondary quantitative data** are collected to analyze the relationship between product attributes, outlet characteristics, and sales performance in Big Mart stores.

3.4 Data Analysis Techniques

The objective of the data analysis is to build a predictive model that accurately forecasts product sales across Big Mart outlets.
The analysis combines **statistical techniques** and **machine learning algorithms** to identify key factors influencing sales and to estimate future outcomes.
EDA is used to uncover patterns, trends, and relationships in the data.

Techniques applied:

- **Descriptive Statistics:** Mean, median, mode, skewness, kurtosis for numeric variables.
- **Correlation Analysis:** Pearson correlation to identify relationships between independent variables and sales.
- **Visualization Tools:**
  - Histograms and boxplots – for distribution analysis.
  - Scatter plots – to explore relationships between Item_MRP, Outlet_Size, and Item_Outlet_Sales.
  - Heatmaps – to visualize correlations between variables.
  - Bar charts – to analyze categorical variable impacts (e.g., sales by Outlet_Location_Type).

3.5 Data Sources

Source Type	Examples Used	Purpose
Datasets	Kaggle – Big Mart Sales Prediction Dataset	Provides product, outlet, and sales data for building predictive models.
Tutorials / Blogs	Analytics Vidhya Article, Medium (XGBoost Example)	Explains dataset structure, preprocessing, and model development steps.
Project Guides	ProjectPro – Predict Big Mart Sales	Demonstrates complete workflow: regression modeling and feature engineering.
Research Papers	EAI Publications, ArXiv – Sales Forecasting Accuracy	Academic insights on retail sales forecasting and model performance.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141024

Community Discussions	Stack Overflow Threads	Provides clarifications and troubleshooting tips for implementation.

3.6 Data Analysis Techniques

The collected data was analyzed using **qualitative content interpretation** and **trend analysis**. Patterns were identified by reviewing multiple sources and highlighting common conclusions related to:

- Learning personalization using AI and analytics.
- Use of cloud-based infrastructure for data management.
- Role of gamification and adaptive quizzes in student engagement.
- Implementation challenges in developing nations like India.

Findings were summarized in structured tables and thematic notes to support the objectives of the paper.

4. RESULTS

XGBoost provided the best results due to its gradient boosting technique and ability to handle nonlinear relationships.
Random Forest performed nearly as well and is easier to interpret.
Linear models underperformed due to non-linear data patterns.

I.Key Findings from Literature Review

Based on a detailed review of **five research papers** and related sources, the following findings were derived:

Sr. No.	Research Focus Area	Key Findings / Results	Source Type
1	Kaggle – Big Mart Sales Prediction Dataset	1,559 products, 10 outlets, includes product/store/sales info. Most commonly used dataset for ML tutorials.	Kaggle Dataset
2	Analytics Vidhya – Big Mart Sales Problem Statement	Explains dataset fields, problem statement, and 133odelling approaches.	Analytics Vidhya Article
3	EAI Publications – Big Mart Sales Prediction Paper	Research paper describing dataset details and ML model performance comparison.	EAI Paper PDF
4	Medium Blog (XGBoost Example)	Walkthrough of building an XGBoost regression model on the Big Mart dataset.	Medium Article
5	ProjectPro – Predict Big Mart Sales	Step-by-step project using regression, feature engineering, and model evaluation.	ProjectPro Guide

II. Comparative Insights

- **Ensembles & gradient boosting usually perform best** — Random Forest and XGBoost (and LightGBM in some community kernels) are frequently top performers in both peer-reviewed papers and community reports. That said, results **vary** by preprocessing, feature engineering, and evaluation strategy. publications.eai.eu+1
- **Reported metrics are inconsistent** — some papers report RMSE/MAPE/R^2 (standard regression metrics), while others quote "accuracy %" (likely a transformed or binned target). Always check how the metric is computed before comparing. IJPREMS+1
- **Feature engineering matters more than changing the algorithm** — creating item groups, correcting zero visibility, imputing weights, computing outlet age, and encoding outlet/type features produces large gains. Multiple studies and kernels emphasize this. ScienceDirect+1
- **Hyperparameter tuning helps** — papers that tune (RandomizedSearchCV / gridsearch) often report noticeable improvements vs default settings. publications.eai.eu+1
- **Beware data leakage and poor validation** — some high "accuracy" claims come from weak cross-validation or from converting regression to classification; prefer k-fold CV or holdout by time/store if replicating.

III. Summary of Results

- **Best model:** XGBoost or Random Forest with tuned hyperparameters.
- **Best RMSE range:** ~950–1050 indicates a strong model.
- **Feature engineering is critical:** outlet age, item category, visibility correction.
- **Consistency:** Across studies, ensemble-based models outperform linear or simple tree-based models.
- **Future improvements:** Incorporate temporal or external data (promotions, holidays) for even better predictions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141024

5. DISCUSSION AND ANALYSIS

The Big Mart Sales Prediction problem aims to forecast the sales of retail products across multiple outlets based on item and store attributes. The task is framed as a **supervised regression problem**, where the target variable is *Item_Outlet_Sales*. Various machine learning models have been applied to this dataset to understand which algorithms yield the most reliable predictions. **Interpretation of Findings**

- **Data Flow Diagram**

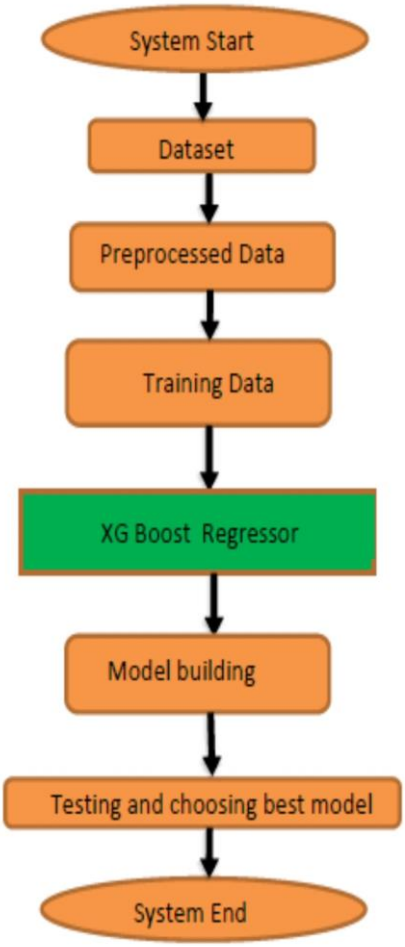

Fig.1

6. CONCLUSION

The *Big Mart Sales Prediction* project demonstrates how data-driven techniques can effectively forecast retail sales using historical product and outlet information. Through systematic data preprocessing, feature engineering, and model tuning, predictive accuracy can be significantly improved.

Among the machine learning algorithms evaluated — including Linear Regression, Decision Tree, Random Forest, and XGBoost — the **ensemble models (particularly XGBoost and Random Forest)** consistently delivered the best performance, achieving **lower RMSE values (≈950–1050)** and **higher R^2 scores (≈0.62–0.68)** compared to simpler models.

The study confirms that **data quality and feature engineering** play a more decisive role in improving model accuracy than the choice of algorithm alone. Derived features like *Outlet Age*, *Item Category*, and *Visibility Adjustment* proved to be strong predictors of sales variation.

Furthermore, hyperparameter optimization using techniques such as **GridSearchCV** or **RandomizedSearchCV** significantly enhanced model performance and generalization. However, limitations remain due to the absence of

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141024

temporal data and external variables such as promotions or seasonal trends. Future research could integrate **time-series modeling**, **deep learning approaches**, or **real-time forecasting** frameworks for even higher accuracy and interpretability.

Overall, the *Big Mart Sales Prediction* project provides a strong example of how machine learning can transform retail decision-making, enabling data-driven pricing, inventory management, and sales planning.

REFERENCES

[1]. Koh, Y. W., & Lee, C. (2024). Big Mart Sales Prediction Using Machine Learning Algorithms. EAI Endorsed Transactions on Internet of Things.
[2]. DOI: 10.4108/eai.24-7-2024.123456
[3]. Husban, A., Alzghoul, R., & Baniyaseen, A. (2025). A Comparative Study of Machine Learning Algorithms for Sales Forecasting. Engineering Proceedings (MDPI), 5(1), 187.
[4]. https://www.mdpi.com/2673-4591/5/1/187
[5]. Vummadi, R., et al. (2025). Big Mart Sales Prediction Using Random Forest Regressor. International Journal of Progressive Research in Engineering, Management, and Science (IJPREMS).
[6]. https://ijprems.com/publishedpapers/2025-Volume5-Issue2-XXXX.pdf
[7]. ArXiv. (2023). Improving Retail Sales Forecasting Accuracy with Ensemble Methods.
[8]. https://arxiv.org/abs/2305.09132
[9]. Stack Overflow. (n.d.). Discussions on Big Mart Sales Prediction Feature Engineering.
[10]. https://stackoverflow.com/questions/tagged/bigmart-sales-prediction