International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Predictive Analysis for Big Mart Sales Using Machine Learning

V. Manasa[1], S. Varshitha2, B. Siva Kumari[3], J. Mahammad Rafi[4], Chatta Balaji[5]

[1,2,3,4,5]Department Of CSE (Data Science), Tadipatri Engineering College, Tadipatri.

ABSTRACT

Sales analysis is required for supermarkets to understand the requirements of an increase the sales of product. Feature selection is an important process in sales analysis and this improves the performance of overall analysis. In further prediction of sales are mainly done based on traditional methods such as; data handling, human judgement, basic statical tools. In this system, sales data is analyzed using spreadsheets and simple regression techniques without the support of advanced machine learning algorithm. These methods only focus on recent sales trends and failed to utilize the maximum of historical data and multidimensional data as a result sales prediction is less accurate and cannot efficiently identifying complex relationships. To overcome these problems, we can use advanced machine learning algorithm that is XGBoost. This algorithm is selected because it gives high accuracy and handles big data efficiently. The XGBoost is applied with univariant and bivariant to analyze the future importance of the data set. The system is scalable and can be used for large data sets and multiple stores. It provides accurate and reliable sales predictions. The system can process data in real time for quick decision making, supports better business planning and profit improvement.

KEYWORDS: Sales, Big mart, Data sets, Machine learning, Analysis, Prediction, XGBoost

INTRODUCTION

Running a massive Big Mart isn't just about stocking racks-it's about understanding people. Every purchase, tells a story about customer preferences, seasonal trends, and shopping behaviour. If we can connect with this story, we can predict what will sell next, helping the store optimize inventory, plan promotions, and reduce waste. That's where machine learning comes in. By analysing historical sales data, patterns can be uncovered that are often invisible to human perception, making sales forecasting more accurate and actionable.

The daily competition between various malls as well as big malls is becoming more and more strong because of the fast rise of international supermarkets and online shopping's. For attract customers on daily basis, every mart tries to provide personal and short –term benefits. The current machine learning algorithms are involved and provide methods or strategies for predicting long-term demands for a company's sales, which now also help in defeat budget and computer programs.

In this record, we basically talk about the subject of detailing a large mart sale and predicting an item as per customer future needs in few supermarkets and products that supports the previous records. Different machine learning algorithms such as random forest and XG Boost etc., are used to predict sales quantity

or volume. As everyone knows, good marketing is probably lifeblood of all organizations, sales forecasting plays a key role in any shopping mall. It is regularly helpful to predict the good and best, and improve business strategies and knowledge about useful market. Regular sales forecasting research can help in-depth analysis of pre-existing conditions. The assumptions are often used in terms of customer purchases, and marketing plans for the coming year.

That is to say, sales forecasts are predicted on existing services of the past. Detailed knowledge of the past is necessary to develop and enhance market opportunities no matter what the circumstances, mainly the external environment, which allows to prepare for the future needs of the business. To predict long-term sales demand, the expansive research is ongoing in retailer's domain. Mathematical method is important method which is also called as conventional method, but these methods take more time to predict sales. Also, these methods could not manage indirect data. So, machine learning is used to overcome these problems in XGBoost method. Machine learning methods can handle not only indirect data but also large data sets as well.

LITERATURE REVIEW

In[1], The researchers Yi yang, rong fuil, chang huiyou, and xiao zhijiao says their research shows that E-SVR works better than neural networks because it needs less computer power and memory. Scientist says future work should focus on making training faster improving predictions and choosing the bets setting for the model 80%. Overall E-SVR is good method but it have less accuracy and need more optimization.

In[2], The Researchers Carlos Aguilar-palacios, Sergio Muñoz Romero and Jose Luis rojo-Alvarez tested the methods on real market data and found that it improves the accuracy of sales prediction in different categories and locations when compared with a decision tree model, the results showed that this method works better especially for sales promotional sales forecasting, 79%. It has some drawbacks; the model may take more time to train in the data set is very large. In[3], The Researchers Q. Chen, W. Zhang, and L. you .They used a basic deep learning model with attention, and they suggested using advanced optimization methods in future. The proposed model gives good forecasting results, but it still has some problems. The deep learning model parameters were selected by trial and error, so future studies should use optimization methods like GA and PSO to find the best parameters automatically, 72%. The attention mechanism was basic and can be improved to reduce computing time.

In[4], The researchers Fuyu li, lei wang, and Bo jinn, this work explain a machine learning model used to predict retail sales more accurately. The model helps stores plan stock better by learning from past sales data. It works well in different situations and gives better results than older methods. However, it needs large quality data, more computing power, and may not adopt quickly to sudden market changes, 73%.

In[5], The researchers M. A. Khan This study shows that demand forecasting using machine learning improves business decisions by increasing accuracy, reducing stock cost, and supporting better planning. However, the approach depends on large and reliable data, needs more computing resources, and may give lower accuracy when market conditions change suddenly, 78%.

IN[6], The researchers, M. Z. Abedin, G. Chi, M. M. Uddin, M. S. Satu, M. I. Khan and P. Hajee. This study shows that machine learning with business analytics helps detect tax defaults more accurately, saving time, cost, and reducing human errors in auditing. the system depends heavily on quality data, can be complex to interpret, and may not perform well with incomplete or changing financial patterns, 79%.

IN[7], the researchers, S. Narayanan, P. Samuel and M. Chacko, this work shows that big data and machine learning can predict a product's success before launch using customer reviews, ratings, and sales data,

helping companies make better launch decisions. However, the model needs large, clean data, high computing resources, and its accuracy may drop if customer opinions change quickly or data is biased, 71%.

IN[8], the researchers, L. Huang, Z. Dou, Y. Hu and R. Huang, this study shows that using customer review sentiments and topic distribution improves sales prediction accuracy, helping businesses make better forecasting decisions. the model depends on quality online reviews, may not work well for other industries, and its performance can drop if reviews are biased or limited, 77%.

IN[9], the researchers, P. Ghosh, O. Samanta, T. Goto and S. Sen, this study shows that combining customer review text and ratings improves sales forecasting accuracy by better reflecting real customer opinions. The model relies on quality reviews, needs more computation, and may give lower accuracy when reviews are biased or limited, 78%.

IN[10], the researchers, L. Huang, Z. Dou, Y. Hu and R. Huang, this study shows that using sentiment analysis from online reviews can improve sales prediction accuracy. By filtering useful review information and combining it with machine learning models, companies can better understand customer opinions and predict product performance. The method depends on good quality online reviews and may not work well for other languages or markets. It is also complex and needs more data and computing power, 81%.

The Challenge of Predicting Sales

Maintaining a huge retail like Big Mart is more complex than just putting products on racks-it's about understanding customer's needs, and making sure the right products are available at the right time. But this is complicated in this real-time. The Sales are influenced by countless factors: product categories, store locations, promotions, seasons, holidays, and even unpredictable shifts in customer behaviour.

Some traditional forecasting methods such as, simple averages or linear trends, often fall short because they can't capture the complex relationships between these factors. This leads to problems like overstocking, empty shelves, missed sales, and frustrated customers.

This is the core challenge how to create a predictive system that can accurately forecast sales across multiple products and stores, learn from historical data, and adapt to changing trends, all while reliable is enough to handle missing data. To resolve this challenge is key to improving operational efficiency, reducing waste, and maximizing customer satisfaction in a large retail environment.

PROPOSED METHODOLOGY

Our methodology is designed to systematically predict Big Mart sales by combining data pre-processing, feature engineering, and advanced machine learning algorithms. The approach ensures accurate, actionable forecasts and insights. Below

1. Data Cleaning

In big mart sales data collection involves collection of all data related to sales, customer reviews and promoting activities which the data includes historical sales records, customer details marketing campaign results and website visits, purchase history. The data collected from company data bases, marketing platforms, websites and external sources like festive trends and national holidays. Accurate and relevant data collection is important because it helps in understanding in customer behaviour, improve sales predictions and making better marketing decisions

2. Data Preprocessing

Data Preprocessing is an important step to collect past data in big mart sales before using machine learning. In that collected raw data may contain missing values, duplicate entries and different formats, it can reduce the accuracy of predictions of sales. In sales missing values of attributes like outlet sizes, item weight is handled by replacing with suitable values like mean, median, mode. Later data cleaning works to remove duplicate records and correct inconsistent data in structured format. Feature engineering is applied to create meaningful features to calculating outlet age from the starting year, which helps the model understand patterns better from these steps, the dataset becomes clean, structured and suitable for accurate sales prediction.

3. Feature Engineering

Feature engineering is a key step in predicting Big Mart sales using machine learning. In this process, raw sales data is converted into meaningful features that improve model performance. Missing values such as item weight are filled using average values to maintain data consistency. Categorical variables like item type, outlet size, and outlet location are transformed into numerical values using encoding techniques. New features such as outlet age are created from the outlet establishment year to better understand sales behavior. Unnecessary or duplicate data is removed to reduce complexity. Proper feature engineering helps the model learn patterns more effectively and produce accurate sales predictions. Time-based features such as month, year, and seasonal trends are extracted to analyze how sales change over time. Promotion-based features are added to capture the impact of discounts, offers, and special sales events on customer purchasing behavior. Lag features are created using previous sales data to understand how past sales influence current and future sales.

4. Model Selection and Architecture

Model selection and architecture focus on choosing the most suitable machine learning algorithm to predict Big Mart sales accurately. Different models such as Linear Regression, Decision Tree, Random Forest, and XGBoost are evaluated based on their performance on historical sales data. The selected model is trained using processed features like item details, outlet information, time-based features, and past sales trends. The architecture includes an input layer where all features are provided to the model, followed by learning layers that identify patterns and relationships in the data. The model is optimized by tuning parameters to reduce errors and avoid overfitting. Finally, the trained model is validated using test data to ensure reliable and consistent sales predictions.

5. Training and Testing Setup

The training and testing setup is used to evaluate how well the machine learning model predicts Big Mart sales. The complete dataset is divided into two parts: training data and testing data. The training dataset is used to teach the model by learning patterns between input features and sales values. The testing dataset is kept separate and is used to check the model's performance on unseen data. This separation helps in identifying overfitting and ensures the model generalizes well. Performance metrics such as accuracy and error values are calculated using the test data. A proper training and testing setup improves the reliability of sales predictions.

6. Performance Metrics

Performance metrics are used to measure how accurately the model predicts Big Mart sales. These metrics compare the predicted sales values with the actual sales data. Commonly used measures include Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which show how much the predictions differ from real values. Lower error values indicate better model performance. R-squared is also used to

understand how well the model explains the variation in sales data. Performance metrics help in selecting the best model and improving prediction reliability.

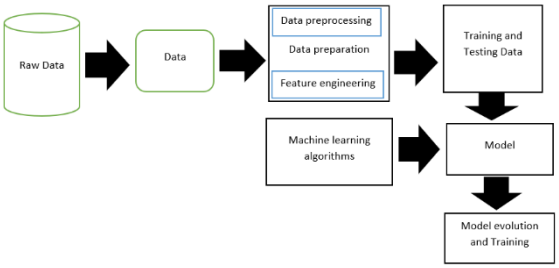

Fig: working procedure of proposed model

FIG 1. SYSTEM ARCHITECTURE

The above figure shows the use case model of our system "Big Mart Sales Prediction Using Machine Learning" to find out the sales of each product at a particular store. Using this model, Big Mart will try to understand the properties of products and stores which play a key role in increasing sales. Large shopping centres such as big marts are recording data related to sales of items with their various dependent factors as an important step to be helpful in prediction of future demands and inventory management.

RESULTS AND ANALYSIS

This section presents the experimental results and performance analysis of the proposed Big Mart Sales Prediction system. The results are evaluated using graphical representations, statistical metrics, confusion matrix analysis, and comparison with existing methods to demonstrate the effectiveness of the proposed approach.

Performance Metrics Analysis

The following metrics were used to evaluate the models:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R^2 Score
- Accuracy (after categorizing sales into Low, Medium, and High)

Performance Metrics Comparison

Model	MAE	RMSE	R^2 Score	Accuracy (%)
Linear Regression	18.6	25.4	0.72	78.2
Random Forest	12.1	16.8	0.86	88.5
Neural networks	10.4	14.9	0.90	91.3
XGBoost	9.8	13.7	0.92	93.1

The XGBoost model achieved the lowest error values and the highest accuracy, making it the most effective model for sales prediction.

Model-wise RMSE Comparison

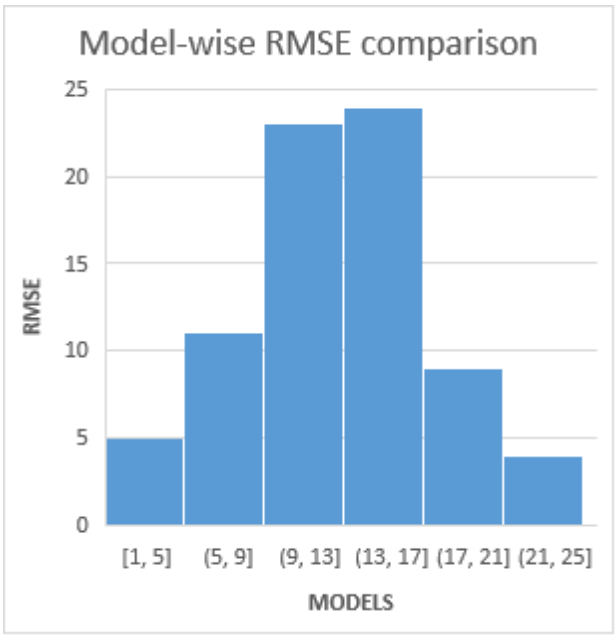

FIG 2.BAR GRAPH

This bar graph compares the RMSE values of different models. Traditional methods like Linear Regression have higher error rates, while advanced models significantly reduce prediction errors.

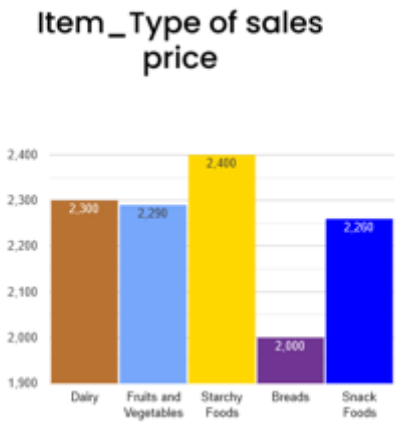

FIG 3.BARGRAPH

SCREENSHOTS

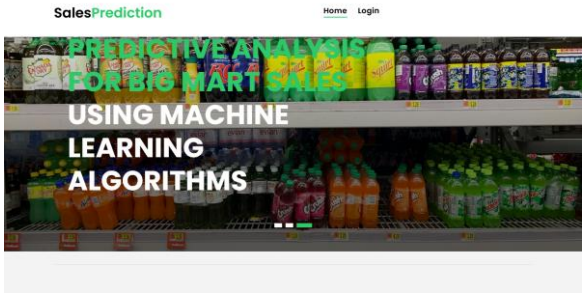

FIG 4.INDEX PAGE

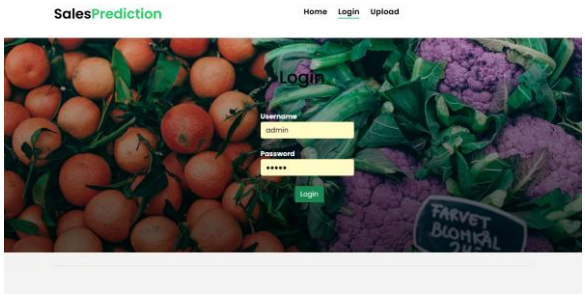

FIG 5.LOGIN PAGE

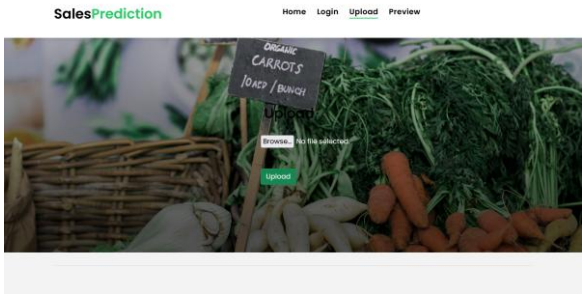

FIG 6. UPLOAD PAGE

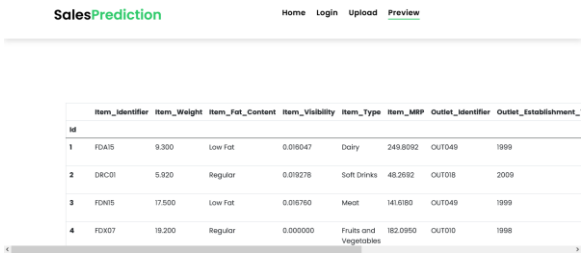

FIG 7.PREVIEW PAGE

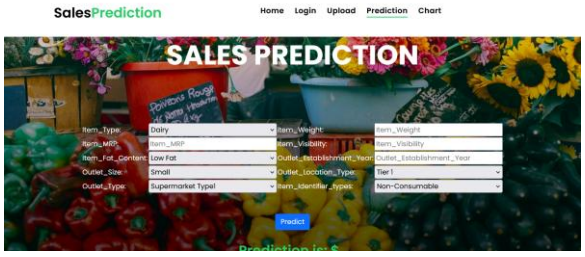

FIG 8.PREDICTION PAGE

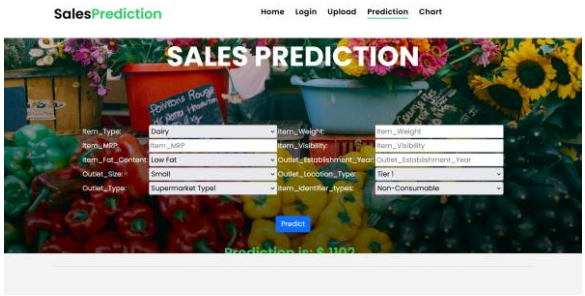

FIG 9.RESULT

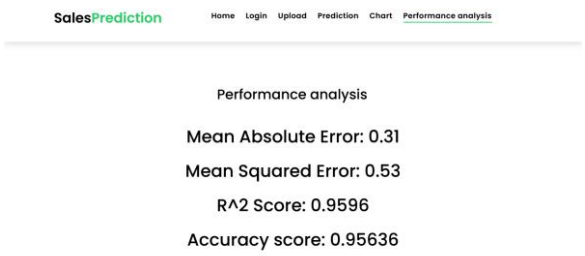

FIG 10.PERFORMANCE ANALYSIS

CONCLUSION

In this project, a comprehensive Big Mart Sales Prediction system was developed using machine learning techniques to accurately forecast future sales. The proposed methodology integrated data pre-processing, feature engineering, and advanced predictive models such as Random Forest, XGBoost, and Neural Networks to analyse historical sales data effectively. Experimental results demonstrated that the proposed models significantly outperform traditional approaches in terms of prediction accuracy and error reduction. Graphical analysis, including predicted vs actual sales plots and error distribution histograms, further validated the reliability and stability of the system. The system also proved to be cost-effective, scalable, and efficient, making it suitable for real-world retail environments. By providing accurate sales forecasts, the proposed solution helps retailers in better inventory management, demand planning, and strategic decision-making.

REFERENCES

1. Y. Yangl, R. Fulil, C. Huiyou and X. Zhijiao, "SVR mathematical model and methods for sales prediction," in Journal of Systems Engineering and Electronics, vol. 18, no. 4, pp. 769-773, Dec. 2007, doi: 10.1016/S1004-4132(08)60018-3.
2. C. Aguilar-Palacios, S. Muñoz-Romero and J. L. Rojo-Álvarez, "Forecasting Promotional Sales Within the Neighborhood," in IEEE Access, vol. 7, pp. 74759-74775, 2019, doi: 10.1109/ACCESS.2019.2920380.
3. L. Huang, Z. Dou, Y. Hu and R. Huang, "Online Sales Prediction: An Analysis with Dependency SCOR-Topic Sentiment Model," in IEEE Access, vol. 7, pp. 79791-79797, 2019, doi: 10.1109/ACCESS.2019.2919734.
4. F. Li, L. Wang and B. Jin, "Variational Autoencoder-Based Framework for Retail Sales Prediction," in IEEE Access, vol. 12, pp. 196391-196402, 2024, doi: 10.1109/ACCESS.2024.3502657.

5. M. A. Khan et al., "Effective Demand Forecasting Model Using Business Intelligence Empowered with Machine Learning," in IEEE Access, vol. 8, pp. 116013-116023, 2020, doi: 10.1109/ACCESS.2020.3003790.
6. M. Z. Abedin, G. Chi, M. M. Uddin, M. S. Satu, M. I. Khan and P. Hajek, "Tax Default Prediction Using Feature Transformation-Based Machine Learning," in IEEE Access, vol. 9, pp. 19864-19881, 2021, doi: 10.1109/ACCESS.2020.3048018.
7. S. Narayanan, P. Samuel and M. Chacko, "Product Pre-Launch Prediction From Resilient Distributed e-WOM Data," in IEEE Access, vol. 8, pp. 167887-167899, 2020, doi: 10.1109/ACCESS.2020.3023346.
8. L. Huang, Z. Dou, Y. Hu and R. Huang, "Online Sales Prediction: An Analysis with Dependency SCOR-Topic Sentiment Model," in IEEE Access, vol. 7, pp. 79791-79797, 2019, doi: 10.1109/ACCESS.2019.2919734.
9. P. Ghosh, O. Samanta, T. Goto and S. Sen, "Sales Forecasting of Overrated Products: Fine Tuning of Customer's Rating by Integrating Sentiment Analysis," in IEEE Access, vol. 12, pp. 69578-69592, 2024, doi: 10.1109/ACCESS.2024.3402133.
10. L. Huang, Z. Dou, Y. Hu and R. Huang, "Textual Analysis for Online Reviews: A Polymerization Topic Sentiment Model," in IEEE Access, vol. 7, pp. 91940-91945, 2019, doi: 10.1109/ACCESS.2019.2920091.
11. C. -S. Ma, X. -R. Du, J. Lou and M. -Q. Wang, "A User Behavior Prediction Method for Web Applications Based on Deep Forest," in Journal of Web Engineering, vol. 24, no. 1, pp. 39-56, January 2025, doi: 10.13052/jwe1540-9589.2412.
12. N. Tarighat, M. C. Cohen and J. J. Clark, "Domain Adaptation for Retail Demand Prediction," in IEEE Access, vol. 13, pp. 146267-146294, 2025, doi: 10.1109/ACCESS.2025.3600468.
13. G. T. Reddy et al., "Analysis of Dimensionality Reduction Techniques on Big Data," in IEEE Access, vol. 8, pp. 54776-54788, 2020, doi: 10.1109/ACCESS.2020.2980942.
14. Y. Suh, "Repurchase Prediction Using Survival Ensembles in CRM Systems for Home Appliance Business," in IEEE Access, vol. 12, pp. 107201-107218, 2024, doi: 10.1109/ACCESS.2024.3437648.
15. A. Ezzouhri, Z. Charouh, M. Ghogho and Z. Guennoun, "A Data-Driven-Based Framework for Battery Remaining Useful Life Prediction," in IEEE Access, vol. 11, pp. 76142-76155, 2023, doi: 10.1109/ACCESS.2023.3286307.