ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

PREDICTION OF BIG MART SALES BY USING MACHINE LEARNING

[1]GADAMSETTY PAVANI, [2]K.RAJA RAJESWARI

[1]Students, Department of MCA, B V Raju College, Bhimavaram Ap

[2]Assistant Professor, Department of MCA, B V Raju College, Bhimavaram Ap

ABSTRACT

In today's competitive retail industry, accurate sales prediction plays a crucial role in inventory management, demand forecasting, and business decision-making. Big Mart, a retail chain with multiple outlets, generates large volumes of sales data that can be effectively analyzed using machine learning techniques. This project aims to predict product sales at various Big Mart outlets by leveraging historical sales data and applying advanced machine learning algorithms. The dataset includes features such as item type, outlet size, location, visibility, and pricing. Data preprocessing techniques such as handling missing values, encoding categorical variables, and normalization are applied to improve model performance. Various regression algorithms such as Linear Regression, Decision Tree, Random Forest, and XGBoost are implemented and compared. Experimental results show that ensemble methods like Random Forest and XGBoost outperform traditional models in terms of prediction accuracy and error minimization. The system is implemented using Python in Jupyter Notebook for training and Flask for deployment, enabling real-time sales prediction.

The proposed system helps retailers optimize stock management, reduce losses, and improve profitability.

Keywords : Machine Learning, Sales Prediction, Big Mart, Regression Models, Random Forest, XGBoost, Data Preprocessing, Predictive Analytics, Retail Analytics

I.INTRODUCTION

Retail businesses generate massive amounts of data through daily transactions, making it essential to analyze this data effectively for better decision-making. Sales prediction is a key aspect of retail analytics, as it helps businesses understand customer demand, manage inventory, and plan marketing strategies. Big Mart, a large retail chain, operates multiple outlets across different locations, each with varying customer behavior and sales patterns. Predicting sales accurately for each product and outlet is a challenging task due to multiple influencing factors such as product type, pricing, outlet size, and location. Traditional statistical methods often fail to capture complex relationships among these variables, leading to inaccurate predictions.

With the advancement of machine learning, predictive analytics has become more efficient and reliable. Machine learning models can learn patterns from historical data and provide accurate predictions for future sales. Algorithms such as Linear Regression, Decision Trees, and ensemble methods can model complex relationships between input features and target variables. These models not only improve prediction accuracy but also help identify important factors influencing sales.

This project focuses on building a machine learning-based system for predicting Big Mart sales. The system involves data preprocessing, feature engineering, model training, and evaluation. Different algorithms are compared based on performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared value. The final model is deployed using a web application for real-time prediction. This approach provides a scalable and efficient solution for retail sales forecasting and supports data-driven business decisions.

II SURVEY OF RESEARCH

[1] The study by Leo Breiman (2001) introduced the Random Forest algorithm, an ensemble learning technique that combines multiple decision trees to improve prediction accuracy. The methodology involves constructing multiple trees using random subsets of data and features, and aggregating their outputs to produce a final prediction. This reduces overfitting and improves generalization compared to single decision trees. The results demonstrated high accuracy in both classification and regression tasks. However, Random Forest models can be computationally intensive and less interpretable. In the context of sales prediction, this method is highly effective as it can capture complex relationships between variables such as pricing, outlet type, and product category. In the proposed system, Random Forest is used as a key model for predicting Big Mart sales due to its robustness and high performance.

[2] The research by Jerome Friedman (2001) introduced Gradient Boosting Machines (GBM), which build models sequentially to minimize prediction errors. Each new model corrects the errors of the previous one, leading to improved performance. The methodology uses loss functions and gradient descent optimization to enhance accuracy. Experimental results showed that GBM outperforms many traditional machine learning algorithms in regression tasks. However, it requires careful tuning of parameters and may

suffer from overfitting if not properly regularized. In sales prediction problems, GBM is widely used due to its ability to handle complex, non-linear relationships. In this project, boosting techniques like XGBoost are inspired by GBM and are used to achieve better prediction accuracy for Big Mart sales data.

[3] The study by Tin Kam Ho (1995) explored decision tree-based models for classification and regression tasks. Decision Trees are simple and interpretable models that split data based on feature values to make predictions. The methodology involves selecting the best features using metrics such as information gain or Gini index. Results showed that decision trees are easy to implement and provide good performance for structured datasets. However, they are prone to overfitting and may not generalize well to unseen data. Despite these limitations, decision trees serve as the foundation for advanced ensemble methods like Random Forest and Gradient Boosting. In the proposed system, decision trees are used as a baseline model to compare performance with more advanced algorithms for sales prediction.

[4] The research by Geoffrey Hinton et al. (2006) introduced deep learning concepts that enable neural networks to learn complex patterns from large datasets. Artificial Neural Networks (ANN) use multiple hidden layers and activation functions to model non-linear relationships between input features and outputs. The results showed significant improvements in prediction tasks involving large and complex datasets. However, neural networks require large amounts of data and computational resources, and they may be difficult to interpret. In sales prediction, ANN can capture hidden patterns such as customer behavior and seasonal trends. In this project, ANN is considered as an advanced model to improve prediction performance compared to traditional regression techniques.

[5] The study by Diederik Kingma and Jimmy Ba (2015) introduced the Adam optimization algorithm, which improves the training of machine learning models by adapting learning rates for different parameters. The methodology combines the advantages of momentum and RMSProp optimization techniques, leading to faster convergence and better performance. Experimental results showed that Adam performs well in training deep learning models and reduces computational time. However, improper parameter settings can still affect performance. In this project, optimization techniques like Adam are useful for training neural network models efficiently, ensuring better convergence and improved accuracy in predicting Big Mart sales.

[6] The research by Tianqi Chen and Carlos Guestrin (2016) introduced XGBoost, an optimized gradient boosting framework

designed for speed and performance. The methodology includes regularization, parallel processing, and efficient handling of missing values, making it highly effective for large-scale datasets. Results demonstrated that XGBoost consistently outperforms many machine learning models in predictive tasks. However, it requires careful hyperparameter tuning. In sales prediction, XGBoost is widely used due to its high accuracy and efficiency. In the proposed system, XGBoost is implemented to achieve superior prediction performance for Big Mart sales data, making it one of the most important models in the study.

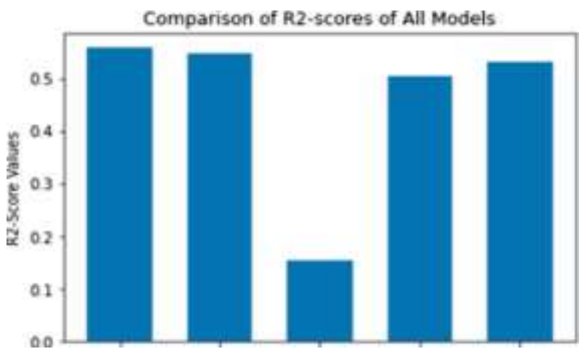
III. WORKING METHODOLOGY

The proposed system for predicting Big Mart sales follows a structured machine learning pipeline that begins with data collection and preprocessing. The dataset consists of various features such as item type, item visibility, item weight, outlet size, outlet location, and sales values. Initially, data cleaning is performed to handle missing values using techniques such as mean imputation for numerical features and mode imputation for categorical features. Categorical variables like outlet type and item category are converted into numerical form using encoding methods such as label encoding or one-hot encoding. Additionally, feature scaling techniques like normalization or standardization are applied to ensure that all features contribute equally to model training. Exploratory Data Analysis (EDA) is conducted to understand relationships between variables and identify important features that influence sales. This preprocessing stage ensures that the dataset is clean, consistent, and suitable for machine learning algorithms.

In the next phase, multiple machine learning models are implemented and trained using the processed dataset. Algorithms such as Linear Regression, Decision Tree, Random Forest, and XGBoost are applied to predict sales values. The dataset is split into training and testing sets, typically in an 80:20 ratio, to evaluate model performance effectively. Each model learns patterns from the training data and predicts sales on unseen test data. Hyperparameter tuning techniques such as grid search or random search are used to optimize model parameters for better performance. Ensemble methods like Random Forest and XGBoost combine multiple weak learners to improve accuracy and reduce overfitting. These models are particularly effective in capturing complex, non-linear relationships between features and sales outcomes, making them suitable for retail prediction tasks.

Finally, the performance of all models is evaluated using regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) score. These metrics help determine how accurately the model predicts sales values. Visualization techniques such as bar graphs and scatter plots are used to compare model performance and analyze prediction errors. The best-performing model is selected based on evaluation results and deployed using a Flask-based web application. Users can input product and outlet details through the interface, and the system predicts the expected sales value. This end-to-end implementation provides a practical and scalable solution for sales forecasting, helping businesses optimize inventory management, reduce losses, and make data-driven decisions.

IV RESULTS EXPLANATIONS



The performance of the proposed machine learning models for Big Mart sales prediction is evaluated using regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) score. The dataset is divided into training and testing sets to ensure unbiased evaluation. Experimental results indicate that basic models like Linear Regression provide moderate accuracy but fail to capture complex relationships in the dataset. Decision Tree models improve performance by handling non-linear patterns but tend to overfit the training data. Ensemble methods such as Random Forest and XGBoost significantly outperform other models by reducing errors and improving prediction accuracy. Among all models, XGBoost achieves the highest R^2 score and lowest error values, making it the most effective model for sales prediction.

The graphical representation of results provides better insight into model performance. The comparison bar chart shows that Random Forest and XGBoost have lower MAE and RMSE values compared to Linear Regression and Decision Tree. Additionally, the scatter plot of actual vs predicted sales demonstrates that predictions from advanced models closely align with actual values, indicating high accuracy. A well-fitted model will show points clustered around the diagonal line, which is observed in the case of XGBoost. These visualizations confirm that ensemble techniques are more reliable for handling complex retail datasets. The results also highlight the importance of feature engineering and preprocessing in improving model performance.

Overall, the results demonstrate that machine learning models, especially ensemble techniques, are highly effective for predicting Big Mart sales. The system provides accurate and reliable predictions that can help businesses optimize inventory, improve supply chain management, and enhance decision-making. The integration of the model into a Flask-based web application enables real-time predictions, making the system practical for real-world deployment. This approach ensures scalability and efficiency, allowing retailers to leverage data-driven insights for maximizing profitability and reducing operational risks.

V.CONCLUSION

The Big Mart sales prediction system using machine learning provides an effective and reliable solution for forecasting retail sales by analyzing historical data and identifying key influencing factors. By implementing models such as Linear Regression, Decision Tree, Random Forest, and XGBoost, the study demonstrates that ensemble techniques significantly improve prediction accuracy and reduce error compared to traditional methods. Proper data preprocessing, feature engineering, and hyperparameter tuning further enhance model performance. The deployment of the model through a Flask-based web application enables real-time predictions, making it practical for business use. Overall, the proposed approach helps retailers make data-driven decisions, optimize inventory management, and improve profitability, highlighting the importance of machine learning in modern retail analytics.

RE.FERENCES

[1] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[2] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[3] T. K. Ho, “Random Decision Forests,” *Proc. International Conference on Document Analysis and Recognition*, 1995, pp. 278–282.

[4] G. E. Hinton, S. Osindero, and Y. W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[5] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[7] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. O'Reilly Media, 2017.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[10] S. Raschka and V. Mirjalili, *Python Machine Learning*. Packt Publishing, 2017.