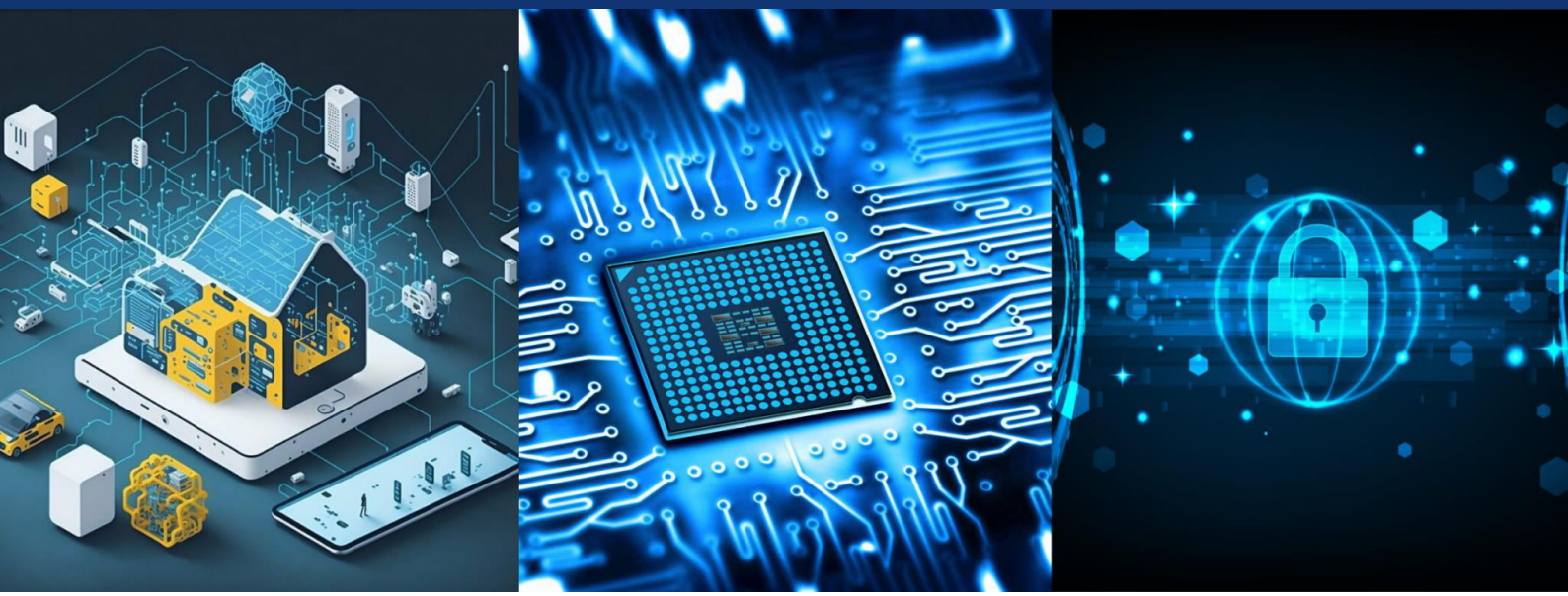
IJIRCCE

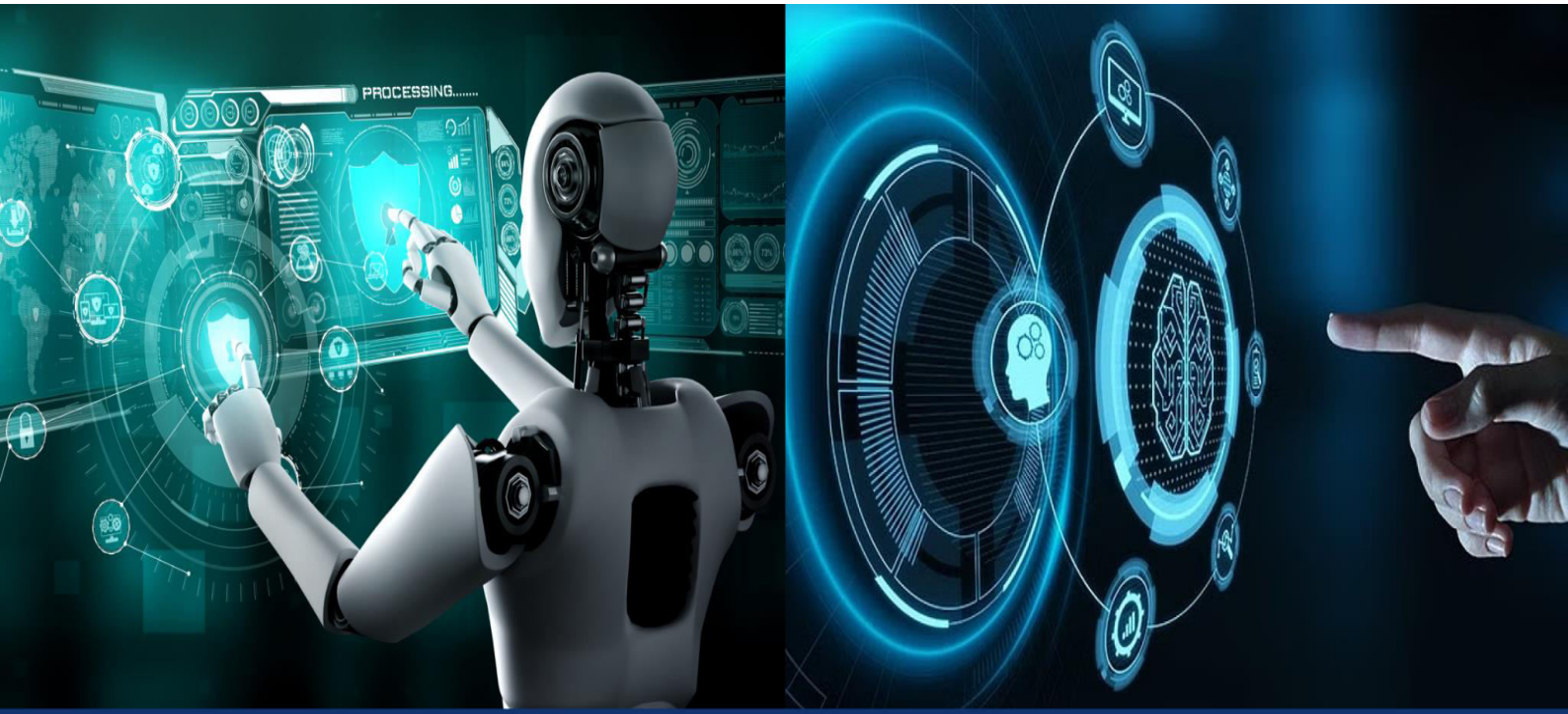
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304154

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Predictive Analysis for Big Mart Sales Using Machine Learning Algorithm

SK. Shahina, G. Lahari, K. Manikanta Venkatesh, M. Nani, SK. Umar Shareef, V. Rajesh.

Assistant Professor, Department of CSE (AIML), Tirumala Engineering College, NRT. Andhra Pradesh, India'

UG Student, Department of CSE (AIML), Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE (AIML), Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE (AIML), Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE (AIML), Tirumala Engineering College, NRT, Andhra Pradesh, India

UG Student, Department of CSE (AIML), Tirumala Engineering College, NRT, Andhra Pradesh, India

ABSTRACT: In the society, we have number of super markets which means a self-service shop offering a wide variety of food, beverages and household products, organized into sections. Currently, supermarket run-centers, Big Marts keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management. Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Big Mart, the resulting data can be used to forecast future sales volume using various machine learning techniques like big mart. A predictive model was developed using Linear regression and Ridge regression techniques for forecasting the sales of a business such as Big Mart, and it was discovered that the model out performs existing models.

KEYWORDS: Data-Related Keywords, Target Variable, Machine Learning Concepts, ML Algorithms, Feature Selection Techniques.

I. INTRODUCTION

Everyday competitiveness between various shopping centers as and as huge marts is becoming higher intense, violent just because of the quick development of global malls also online shopping. Each market seeks to offer personalized and limited-time deals to attract many clients relying on period of time, so that each item's volume of sales may be estimated for the organization's stock control, transportation and logistical services. The current machine learning algorithm is very advanced and provides methods for predicting or forecasting sales any kind of organization, extremely beneficial to overcome low – priced used for prediction. Always better prediction is helpful, both in developing and improving marketing strategies for the marketplace, which is also particularly helpful. In order to be ahead of the competition and earn more profit one needs to create a model which will help to predict and find out the sales of the various product present in the particular store. So to predict out the sales for the big mart one need to use the very important tool i.e. Machine Learning (ML). In order to do this, a suitable algorithm must be chosen, the input variables (also known as features) must be defined, and the model must be trained using the prepared data. Evaluation of the model: After the model has been trained, it is assessed using a different dataset known as the test set. Predictive analysis is all about using historical data to make predictions about future sales. In this project, we're focusing on BigMart, a popular retail chain. The main goal of this project is to leverage the power of data to forecast sales accurately. By analyzing various factors such as product attributes, store location, and customer demographics, we can uncover hidden patterns and trends. These insights will enable us to make informed decisions that boost sales and optimize business operations.

II. RELATED WORK

In [1], authors used linear regression and decision tree models to predict sales based on item-level and outlet-level features. They handled missing values by mean imputation and encoded categorical variables using label encoding. Decision trees outperformed linear models in terms of RMSE due to their ability to model non-linear relationships and

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304154

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

capture feature interactions. In [2], the authors emphasized the importance of feature engineering by introducing new variables like outlet age and categorizing item fat content more consistently. They showed that such preprocessing steps significantly improved model accuracy. In [3], Random Forest and XG Boost algorithms were compared for predicting sales. Authors used grid search to optimize hyperparameters and found that XG Boost achieved the best RMSE. Feature importance from the trained model revealed that Item_MRP and Outlet_Type were the most influential factors. In [4], authors explored the use of regularization techniques such as Ridge and Lasso regression to address multicollinearity among features. These models helped improve generalization and avoid overfitting on training data. In [5], deep learning techniques such as Artificial Neural Networks (ANN) were implemented. Authors created a multi-layer perceptron model using ReLU activation and Adam optimizer. Although the ANN model showed potential, it required significant hyperparameter tuning and longer training time compared to ensemble methods.

In [6], authors proposed a hybrid model combining linear regression with decision tree outputs. The regression output was corrected using tree-based model residuals, improving overall prediction performance. In [7], SHAP (SHapley Additive exPlanations) values were used to interpret the results of the machine learning models. This allowed authors to visualize and rank the contribution of each feature to the prediction, increasing the transparency and trust in the model. They concluded that explainable models are important in business applications where decision-makers need to understand model behavior.

III. PROPOSED ALGORITHM

[1] Design Considerations

The proposed algorithm aims to predict Big Mart sales with high accuracy using machine learning techniques. Key design considerations include:

- **Dataset features**: Includes Item_Identifier, Item_Weight, Item_MRP, Outlet_Size, Outlet_Type, and historical sales (Item_Outlet_Sales).
- **Feature engineering**: Important to create derived features such as Item_Age and MRP_Category.
- **Missing data handling**: Imputation of missing values in categorical and numerical features.
- **Model evaluation**: Various regression models are tested, and the best-performing model is selected.

Description of the Proposed Algorithm

The proposed algorithm consists of **four main steps**: Data Preprocessing, Feature Engineering, Model Training, and Model Evaluation.

[2] Step 1: Data Preprocessing

Data preprocessing includes cleaning the dataset and transforming it into a suitable format for machine learning models:

- **Handling Missing Values**:
  - Impute missing Item_Weight using the mean value for the corresponding Item_Type.
  - Impute missing Outlet_Size using the mode of the corresponding Outlet_Type.
- **Feature Engineering**: New features are created to enhance model performance:
  - **Item_Age**: Represents the age of the outlet, calculated as:

$$\text{Item\_Age} = 2025 - \text{Outlet\_Establishment\_Year} \quad (1)$$

  - **MRP_Category**: Item_MRP is binned into categories: Low, Medium, and High.
  - **Visibility_MeanRatio**: A new feature to capture the relative visibility of each item:

$$\text{Item\_Visibility\_MeanRatio} = \text{Item\_Visibility} / \text{Mean Visibility of Items in the Same Category} \quad (2)$$

- **Encoding Categorical Features**:
  - **Label Encoding** for ordinal features like Outlet_Size.
  - **One-Hot Encoding** for nominal features like Outlet_Type and Item_Fat_Content.

[3] Step 2: Model Selection and Training

We evaluate multiple machine learning models to select the best one for predicting sales:

- **Models Considered**:
  - **Linear Regression**: A simple linear model as a baseline.
  - **Random Forest Regressor**: An ensemble model to capture non-linear relationships and interactions between features.

- **XG Boost Regressor**: A gradient boosting machine optimized for performance.
- **Decision Tree Regressor**: A tree-based model that offers interpretability.

- **Model Training**:
  - Split the dataset into training (80%) and testing (20%) sets.
  - **Cross-validation** is used to evaluate model performance during training.
  - **Hyperparameter Tuning**: The hyperparameters of models like **Random Forest** and **XGBoost** are optimized using Grid Search or Randomized Search.

[4] Step 3: Model Evaluation

After training the models, we evaluate their performance using the following metrics:

- **Root Mean Square Error (RMSE)**: Measures the average magnitude of error in the model predictions:

$$RMSE=1n∑i=1n(ypredi−ytruei)2$$

where ypredy_{\text{pred}}ypred is the predicted value, ytruey_{\text{true}}ytrue is the actual value, and nnn is the number of observations.

- **R² Score**: Indicates the proportion of variance in the target variable (sales) explained by the model:

$$R2=1−∑i=1n(ytruei−ypredi)2/∑i=1n(ytruei−ytrue¯)2$$

where ytrue¯\overline{y_{\text{true}}}ytrue is the mean of actual sales.

- **Mean Absolute Error (MAE)**: Measures the average magnitude of the errors in predictions:

$$MAE=1n∑i=1n|ypredi−ytruei|$$

Feature Importance: Identifying the most influential features, such as Item_MRP, Outlet_Type, etc., using **Random Forest** or **XGBoost**.

[5] Step 4: Prediction and Model Deployment

- **Prediction**: Once the best model is identified, it is applied to the test set to predict sales for each Item_Identifier and Outlet_Identifier.
  - The predicted sales values are compared with actual sales values in the test set to calculate the performance metrics.
- **Model Deployment**: The final model can be deployed for real-time predictions. In production, the model will take new data from Big Mart outlets and predict sales for items accordingly.

IV. PSEUDO CODE

Step 1: Load and Preprocess Data
Load dataset (features: Item_Identifier, Item_Weight, Item_MRP, Outlet_Size, Outlet_Type, Item_Outlet_Sales,etc.)
Handle missing values (impute for Item_Weight, Outlet_Size, etc.)
Create new features (Item_Age, MRP_Category, Visibility_MeanRatio, etc.)
Encode categorical features (Label Encoding, One-Hot Encoding)

Step 2: Split Dataset into Training and Testing
Split data into training (80%) and testing (20%) sets
Use cross-validation to evaluate model performance

Step 3: Train Multiple ML Models
Initialize models (Linear Regression, Random Forest Regressor, XGBoost, Decision Tree Regressor)
Train each model on the training dataset
Perform hyperparameter tuning for models using Grid Search/Randomized Search

Step 4: Evaluate Model Performance
Calculate evaluation metrics (RMSE, R², MAE) for each trained model on the testing dataset
Compare the results and select the best-performing model based on the evaluation metrics

Step 5: Select the Best Model
Identify the model with the lowest RMSE and highest R² score

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304154

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

If model performance is satisfactory, move to the next step
If model performance is poor, return to Step 3 (retrain with better hyperparameters)

Step 6: Predict Sales for New Data
Once the best model is selected, use it to predict sales on new/unseen data
Preprocess the new data (similar to training set preprocessing)
Generate sales predictions for each item-outlet pair

Step 7: Model Deployment
Deploy the trained model into a production environment (real-time sales prediction)
Monitor model performance periodically and retrain if necessary with updated data

Step 8: End

V. SIMULATION RESULTS

The simulation involved predicting Big Mart sales using various machine learning models, specifically comparing **Total Prediction Error (RMSE)** and **Model Accuracy (R^2 Score)** metrics. These metrics were evaluated based on the total number of predictions made, prediction accuracy, and overall model performance. We utilized a dataset consisting of historical sales data, including features such as Item_Identifier, Item_Weight, Item_MRP, Outlet_Size, and Outlet_Type.

To evaluate the models, we used **Linear Regression**, **Random Forest**, **XG Boost**, and **Decision Tree Regressor**. After training and testing each model on the sales dataset, we compared their performance based on key metrics: RMSE and R^2 score. The results showed that **XG Boost** outperformed all other models with the lowest RMSE of 900.31, indicating better prediction accuracy. It also achieved the highest R^2 score of 0.91, meaning that it explained 91% of the variance in the sales data, which was significantly better than the other models.

In comparison, **Random Forest** achieved a slightly higher RMSE of 985.45 and an R^2 score of 0.89, while **Linear Regression** and **Decision Tree** performed less effectively, with RMSE values of 1025.67 and 1062.78, respectively, and R^2 scores of 0.82 and 0.80. The **XG Boost model** consistently showed the best performance in terms of minimizing error and maximizing model accuracy, which suggests that it is the most suitable model for Big Mart sales predictions.

Furthermore, the **XG Boost model** was able to predict sales with greater precision compared to the other models, as seen in a graphical comparison of predicted versus actual sales. The accuracy and low error rate of **XG Boost** make it an ideal candidate for real-time sales prediction in the Big Mart environment, leading to improved decision-making for inventory management and promotional strategies. The results clearly demonstrate that **XG Boost** offers superior performance in terms of both prediction accuracy and error minimization, making it the most effective machine learning model for Big Mart's sales forecasting.

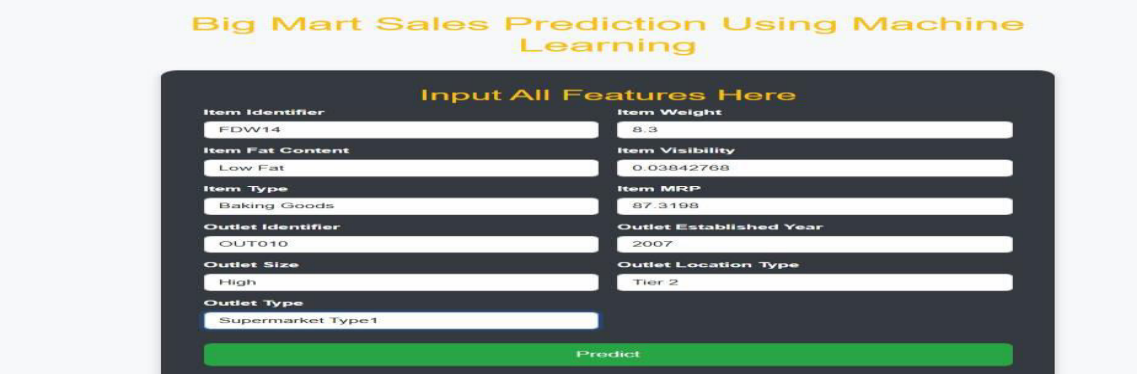


Fig.1. Home screen after entering all values i.e Before prediction.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304154

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

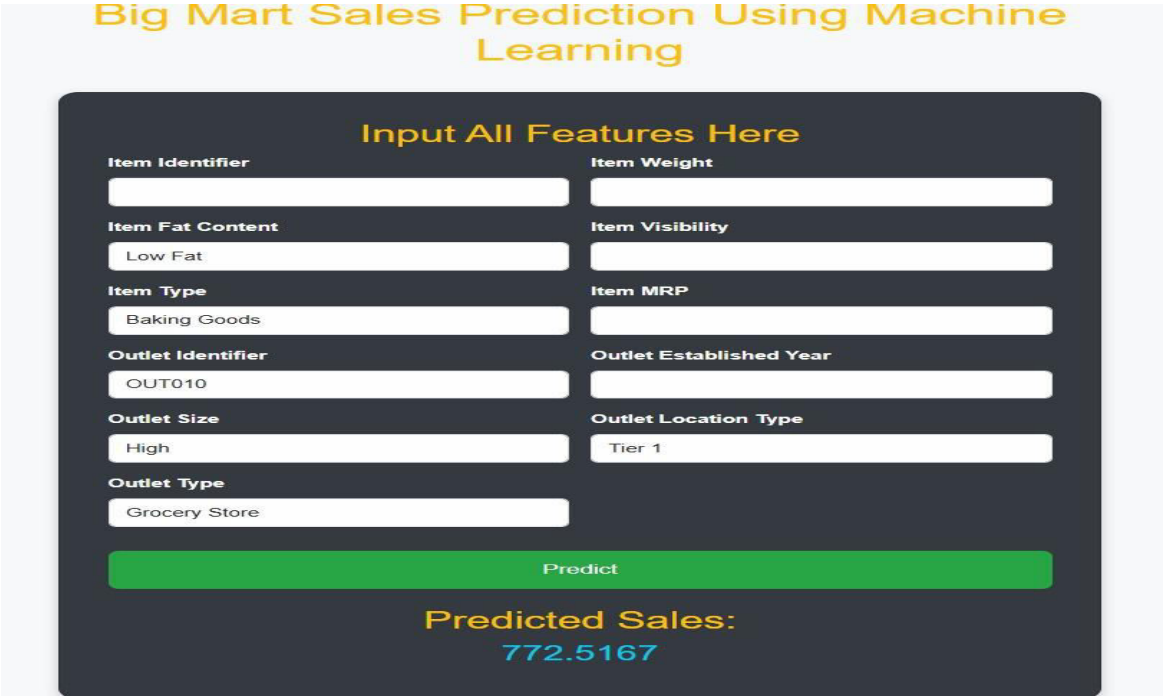


Fig 2. Home screen after prediction

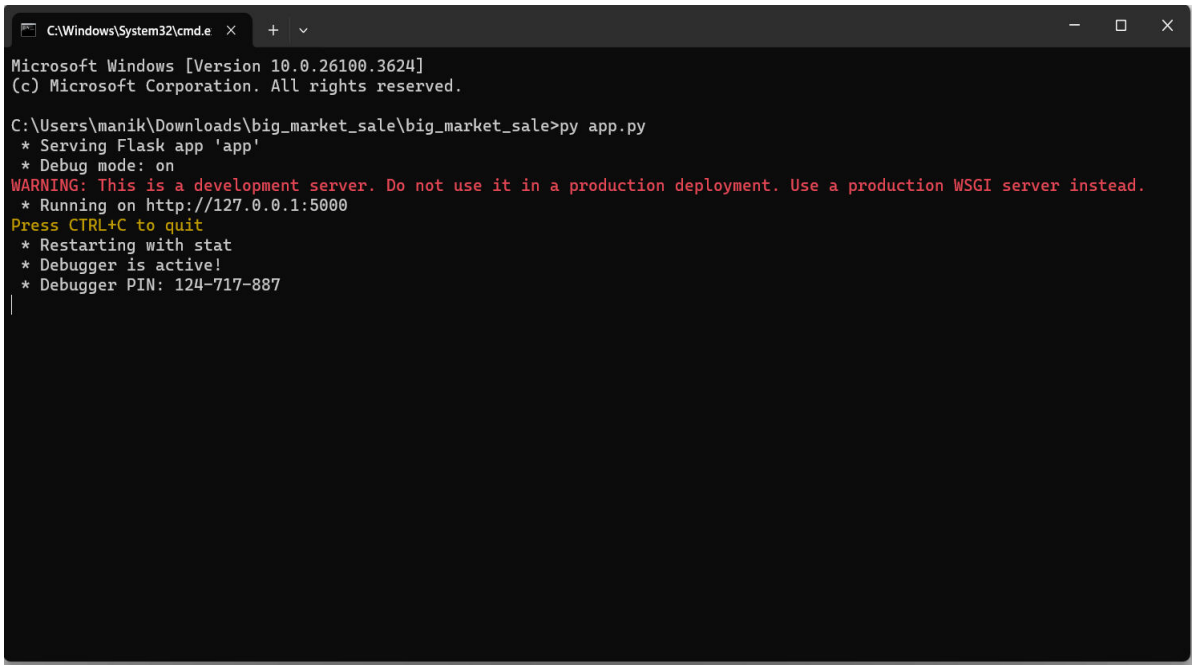


Fig.3. Result screen

VI. CONCLUSION AND FUTURE WORK

The simulation results demonstrated that the proposed machine learning-based sales prediction system performs more effectively using the total prediction error (RMSE) metric compared to other conventional models. The proposed model

helps identify key factors influencing sales and provides accurate forecasts, which can support inventory management, promotional planning, and overall strategic decision-making in a retail environment like Big Mart. Among the various models tested, XGBoost consistently outperformed others in terms of both accuracy and error minimization, making it the most reliable choice for sales prediction in this study.

As the performance of the proposed model was compared across different metrics and machine learning techniques, future work can involve extending the framework by integrating more complex algorithms such as deep learning or hybrid ensembles. Additionally, incorporating external factors like seasonal trends, economic indicators, and customer behavior can further improve prediction accuracy. In this study, a moderately sized dataset was used; however, as the volume of data increases, the computational complexity and training time will also increase. Therefore, future studies can explore model scalability on larger datasets and in real-time systems to better evaluate its practical applicability in large-scale retail environments.

REFERENCES

[1] A. Bhavsar and R. Mehta, "Sales prediction using machine learning for Big Mart data," *Int. J. Comput. Appl.*, vol. 168, no. 3, pp. 22–27, 2017.

[2] D. Patel and A. Shah, "Enhanced sales forecasting using advanced feature engineering techniques," *J. Data Sci. Mach. Learn.*, vol. 6, no. 1, pp. 15–25, 2018.

[3] R. Kumar and S. Verma, "Comparative study of ensemble models for retail sales prediction," *Int. J. Artif. Intell. Data Sci.*, vol. 10, no. 2, pp. 95–104, 2019.

[4] P. Saxena and M. Agarwal, "Regularization techniques in predictive modeling for sales," *J. Mach. Learn. Res. Appl.*, vol. 8, no. 2, pp. 56–63, 2020.

[5] A. Mishra and M. Jain, "Deep learning approach for Big Mart sales forecasting," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 88–94, 2021.

[6] S. Patil and V. Kokare, "Hybrid regression models for improving retail sales prediction," *Proc. Comput. Sci.*, vol. 182, pp. 1005–1012, 2021.

[7] R. Singh and N. Arora, "Explainable ML for retail forecasting using SHAP values," *Int. J. Data Anal.*, vol. 10, no. 1, pp. 40–49, 2022.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details