

## **Machine Learning-Based Price Prediction for Intelligent Retail Management**

**Dr. Prashant Kumar Shrivastava**

PhD CSE, Associate Professor Sanjeev Agrawal Global Educational (SAGE) University Bhopal SOCT

**ABSTRACT:** Retail price prediction is an important component of intelligent retail analytics, supporting effective pricing strategies, revenue optimization, inventory planning, and data-driven decision-making. Because of the complexity and nonlinearity of large-scale retail data, traditional statistical methods may be limited. The Walmart Store Sales Forecasting Dataset is a dataset of 421,570 records with 16 variables, proposed for use in this study to develop a machine-learning predictive price analytics framework. The methodology includes missing-value treatment, duplicate removal, outlier management, one-hot encoding, Min–Max normalization, and SelectKBest feature selection. The dataset was divided into training and testing sets using an 80:20 ratio, and a Random Forest (RF) regression model was developed for prediction. The proposed RF model achieved an R-squared value of 98.96%, RMSE of 0.488, MAE of 0.317, and MSE of 2.386. These results demonstrate the effectiveness of the proposed approach in capturing complex retail sales patterns and producing accurate predictions. The framework can enable intelligent pricing, enhance retail management performance. In addition, the proposed approach can be scaled to process large retail data and gain insights into factors influencing sales performance. It can also help retailers make data-informed pricing decisions and optimize their operations in competitive markets.t)

**KEYWORDS:** Demand Forecasting, Dynamic Pricing, Inventory Management, Price Prediction, Retail Optimization.

### **I. INTRODUCTION**

E-commerce is increasingly necessary to meet changing customer preferences, provide convenient, accessible shopping, expand business reach, reduce operational barriers, and support efficient, competitive, and customer-focused retail management. Retail is a major component of the global economy, connecting businesses with consumers through physical stores, e-commerce platforms, and other distribution channels [1]. Effective retail management requires an accurate understanding of customer demand, product sales, market trends, and competitive pricing [2][3]. Appropriate pricing decisions directly influence revenue, customer satisfaction, inventory turnover, and overall business performance [4]. However, rapidly changing consumer preferences, seasonal demand, competition, and market conditions make accurate price determination increasingly challenging [5][6].

Traditional pricing and demand forecasting approaches largely depend on historical data and predefined statistical relationships. Although these methods can provide useful insights, they often struggle to respond effectively to sudden changes in market conditions, consumer behavior, and competitor prices [7][8]. Accurate demand forecasting is also essential for maintaining appropriate inventory levels, reducing stockouts and overstocking, minimizing operational costs, and improving customer satisfaction. The rapid expansion of e-commerce has further increased the need for flexible pricing strategies capable of responding to continuously changing online market conditions [9][10].

Dynamic pricing has therefore become an important strategy in modern retail, where product prices can be adjusted according to demand, competition, customer behavior, inventory levels, and other market factors. Previous studies have demonstrated that adaptive pricing strategies can improve revenue generation and support better inventory management [11][12]. In particular, the growing availability of retail transaction data provides

opportunities to identify complex relationships among product characteristics, demand patterns, sales history, and pricing factors. Such data-driven approaches can support retailers in making more informed and timely pricing decisions [13][14].

Recently, Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) have emerged as powerful technologies for intelligent retail analytics. ML algorithms can learn complex patterns from historical and market data to predict suitable product prices, while DL techniques can capture nonlinear relationships and temporal patterns in large-scale retail datasets [15][16]. Therefore, a Machine Learning-Based Price Prediction for Intelligent Retail Management approach can provide an adaptive and data-driven solution for accurate price prediction, helping retailers improve pricing decisions, optimize inventory, increase profitability, and remain competitive in dynamic market environments [17].

#### **A. Motivation and Contribution**

The motivation behind this study is the increasing need for accurate and adaptive product price prediction in dynamic retail environments. Traditional pricing approaches may not

effectively capture complex relationships among sales, demand, market conditions, and other influencing factors. Machine learning can analyze large-scale retail data and identify hidden patterns to support reliable and data-driven pricing decisions. Therefore, this study aims to develop an intelligent price prediction framework for improving pricing strategies, inventory management, and overall retail performance. This research offers several key contributions as listed below:

- A machine learning-based predictive price analytics framework is proposed for intelligent retail management.
- The Walmart Store Sales Forecasting Dataset containing 421,570 records and 16 variables is utilized for model development.
- A comprehensive data preprocessing pipeline is applied, including missing-value treatment, duplicate removal, outlier management, one-hot encoding, and Min–Max normalization.
- SelectKBest feature selection is employed to identify the most influential variables for retail sales prediction.
- A Random Forest (RF) regression model is developed to capture complex and nonlinear relationships in retail data.
- The framework provides a data-driven approach for improving pricing decisions, sales analysis, and intelligent retail management.
- The proposed methodology can support scalable and efficient retail decision-making in competitive business environments.

The proposed study is justified by the need for accurate and reliable retail price prediction to support effective pricing and data-driven business decisions. Traditional approaches may not adequately capture the complex and nonlinear patterns present in large-scale retail sales data. The novelty of this work lies in integrating comprehensive preprocessing, Select Best feature selection, and Random Forest regression into a unified predictive price analytics framework. The proposed methodology identifies influential retail attributes and enhances the efficiency of the prediction process. It provides a scalable and practical solution for intelligent retail management and supports improved pricing strategies and operational decision-making.

#### *B. Paper Structure and Organization*

The paper is organized as follows: Section II presents related studies, Section III describes the dataset and methodology, Section IV discusses the experimental results and comparison, and Section V concludes the study with key findings and future research directions.

## **II. LITERATURE REVIEW**

The existing literature on Predictive Price Analytics for Intelligent Retail Management was thoroughly reviewed to understand research findings, trends, and gaps in prior studies that would inform the development of the proposed study.

Preethika et al. (2026) introduce a dynamic pricing setup built around machine learning. It forecasts the best prices for items by drawing on past sales records, competitor prices and

price-sensitivity calculations. The approach uses a Random Forest model for regression. Training happens on features crafted from demand factors and delayed sales metrics, plus elasticity estimates, to figure out prices that adjust over time for store goods. Testing showed the model hitting an  $R^2$  value of 0.93. Mean absolute error came in at 27.66. Against past pricing strategies the ML method simulated revenue growth of 165% [18].

Admeur and Zine (2026) propose a hybrid forecasting framework for Moroccan retail diesel prices by combining Seasonal ARIMA with exogenous variables (SARIMAX) and Support Vector Regression (SVR). SARIMAX models the linear time-series structure using gasoline prices as exogenous signals, while SVR learns nonlinear residual corrections. Using bi-weekly diesel and gasoline prices from 2014–2022, the hybrid model improved the coefficient of determination by 14.6% over SARIMAX and 7.6% over SVR, while reducing mean squared error by 39.7% and 26.3%, respectively.[19].

Padila, Yuliyanti and Al Husaini (2025) aim to help retail companies predict product prices more accurately using machine learning instead of estimates. The Random Forest model initially achieved an  $R^2$  of 80%, MAPE of 25%, and RMSE of 1.07%. After feature selection and importance analysis, performance improved to an  $R^2$  of 92%, MAPE of 13%, and RMSE of 0.84%, indicating a 15.3% increase in prediction accuracy. The content-based product recommendation system using cosine similarity achieved the highest similarity score of 0.42% [20].

Poongothai et al. (2024) propose an ML-based system that uses the most innovative algorithms such as Long Short-Term Memory and XGBoost to avoid such problems. The use of specific data and algorithms helped improve operational efficiency and reduce MAPE to 6.8% for short-term forecasts. As a result, the system demonstrated an improved inventory turnover ratio of 6.3 and a considerable decrease in the percentage of stockouts and overstock events [21].

Chen et al. (2024) aim to predict retail cigarette sales using neighboring alcohol sales data through Propensity Score Matching (PSM), KMeans++ clustering, and XGBoost. Using cigarette and alcohol sales data from Guangzhou, China, for 2021–2022, the results show that introducing neighboring alcohol sales data with PSM and XGBoost improved modeling accuracy from 82.79% to 91.34%. XGBoost identified category, specification, price range, and neighboring alcohol sales revenue as important features.[22].

Upadhyay et al. (2023) involved the implementation of several machine learning algorithms, including linear regression, decision tree, random forest, gradient boosting, support vector machine, convolutional neural network, and long short-term memory. The models were trained and evaluated using a dataset comprising over 8,000 entries of sales data from different BigMart stores. The results showed that the convolutional neural network and long short-term memory models outperformed the other algorithms in terms of accuracy, with  $R^2$  values of 0.68 and 0.69, respectively [23].

McRae, Nie and Dube, (2022) focus on predicting demand for 13 newly launched yogurt products using 18 machine learning models at market and store levels. Store-level data were enriched with 2016 Canadian census demographic data, and 330 features were engineered covering product, marketing, promotions, store, and neighborhood information. The best market-level Random Forest model achieved an  $R^2$  of 84.0%,

while the store-level LightGBM model achieved 57% on test data [24].

A review of existing studies on Predictive Price Analytics for Intelligent Retail Management was conducted to identify key approaches and research gaps. Table I summarizes the major studies and their contributions.

**TABLE I. RECENT STUDIES ON PREDICTIVE PRICE ANALYTICS FOR INTELLIGENT RETAIL MANAGEMENT**

| Author                                  | Proposed Work                                                                                                           | Dataset                                           | Key Findings                                                                                       | Limitations & Future Work                                                                                           |
|-----------------------------------------|-------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|----------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Preethika et al. (2026)                 | ML-based dynamic pricing using Random Forest with sales, competitor prices, demand, lagged sales, and price elasticity. | Historical retail sales and competitor-price data | $R^2 = 0.93$ , MAE = 27.66; simulated revenue improved by 165%.                                    | Requires real-time validation, larger product categories, and real-world deployment.                                |
| Admeur and Zine (2026)                  | Hybrid SARIMAX + SVR for linear and nonlinear fuel-price forecasting.                                                   | Moroccan diesel and gasoline prices, 2014–2022    | $R^2$ improved by 14.6% over SARIMAX and 7.6% over SVR; MSE also decreased.                        | Focused on fuel prices; future work can include other retail products and market variables.                         |
| Padila, Yuliyanti and Al Husaini (2025) | ML-based product price prediction with feature selection and product recommendation.                                    | Retail product pricing and market-trend data      | RF improved $R^2$ from 80% to 92%, MAPE from 25% to 13%, and RMSE from 1.07 to 0.84.               | Requires larger real-time datasets and additional pricing factors.                                                  |
| Poongothai et al. (2024)                | Sales-demand forecasting using LSTM and XGBoost for inventory decisions.                                                | Retail sales and inventory data                   | MAPE = 6.8%, inventory turnover = 6.3, with reduced stockout and overstock events.                 | Future work can integrate dynamic pricing and real-time market information.                                         |
| Chen et al. (2024)                      | Retail sales prediction using PSM, KMeans++, and XGBoost with related alcohol sales.                                    | Guangzhou cigarette and alcohol sales, 2021–2022  | Accuracy improved from 82.79% to 91.34%.                                                           | Limited product and geographic scope; future work could include broader product coverage, competitors, and pricing. |
| Upadhyay et al. (2023)                  | Comparative ML/DL sales prediction using LR, DT, RF, GB, SVM, CNN, and LSTM.                                            | BigMart sales dataset, 8,000+ records             | CNN and LSTM achieved $R^2 = 0.68$ and 0.69; key factors included weight, location, and promotion. | Performance remains moderate; pricing, competitors, and hybrid models warrant exploration.                          |
| McRae, Nie and Dube (2022)              | ML-based demand prediction using product, marketing, store, and demographic features.                                   | 13 yogurt products with store and census data     | RF achieved $R^2 = 84\%$ at market level; LightGBM achieved 57% at store level.                    | Store-level prediction remains challenging; richer temporal and customer data are needed.                           |

**Research Gaps :** Existing studies demonstrate the effectiveness of machine learning for retail sales forecasting, demand prediction, and dynamic pricing, but many focus on specific products, markets, or limited datasets. Most use individual forecasting or pricing models without comprehensive preprocessing and feature selection for big retail data. Limited attention is also given to Random Forest for predictive price analytics using diverse retail attributes. Therefore, a unified framework is needed to identify influential features and model nonlinear retail sales relationships. To address these gaps, this

study develops an integrated machine learning framework for price analytics to support intelligent retail management.

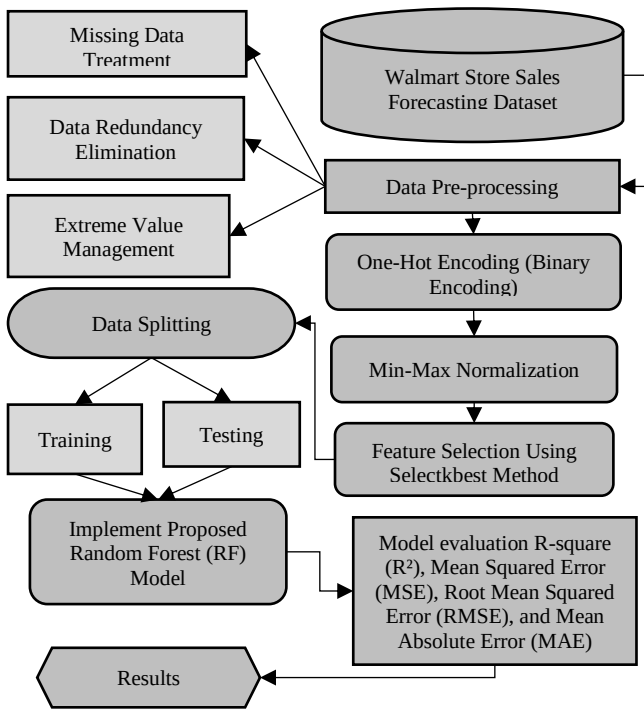


Fig. 1. Proposed Framework for Machine Learning-Based Predictive Price Analytics in Intelligent Retail Management

### III. RESEARCH METHODOLOGY

The proposed methodology comprises data collection, exploratory analysis, preprocessing, feature selection, and model development, using the Walmart Store Sales Forecasting Dataset. Data cleaning involves treatment of missing values, removal of duplicate data, outlier treatment, one-hot encoding, Min–Max normalization, and feature selection using SelectKBest with 80:20 training/validation split. Lastly, the  $R^2$ , MSE, RMSE and MAE are used to evaluate the predictive performance of the Random Forest regression model for intelligent retail price analytics. The proposed flowchart for the intelligent retail management system, using machine learning for predictive price analytics, is presented in Fig. 1. The following section presents a detailed explanation of each step involved in the proposed methodology:

#### A. Data Collection and Exploratory Analysis

The Walmart Store Sales Forecasting Dataset, openly available on Kaggle, was used for this study and contains 421,570 records with 16 variables, including Weekly\_Sales as the target variable for intelligent retail price prediction. Data visualizations such as bar plots and heatmaps were used to examine sales patterns, feature distributions, and correlations among variables:



Fig. 2. Distribution of Weekly Sales

Fig. 2 shows the distribution of Weekly Sales using a histogram with a smooth KDE curve. Most sales values are concentrated around 0.4–1.5 million, while the frequency gradually decreases as sales increase. The long right tail indicates a right-skewed distribution, with a small number of observations having very high weekly sales.

Fig. 3: The correlation matrix heatmap shows the relationships between the variables in the dataset. Weekly\_Sales has very weak correlations with the other variables, while CPI and Unemployment have the strongest negative correlation of -0.30. Overall, the variables have low correlations, indicating limited multicollinearity and that each variable can provide useful information for the prediction model.

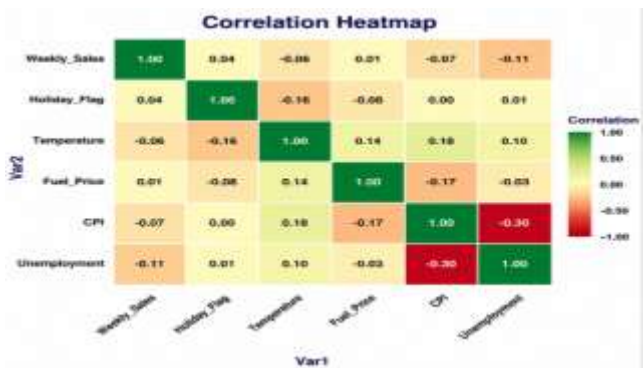


Fig. 3. Correlation Matrix Heatmap

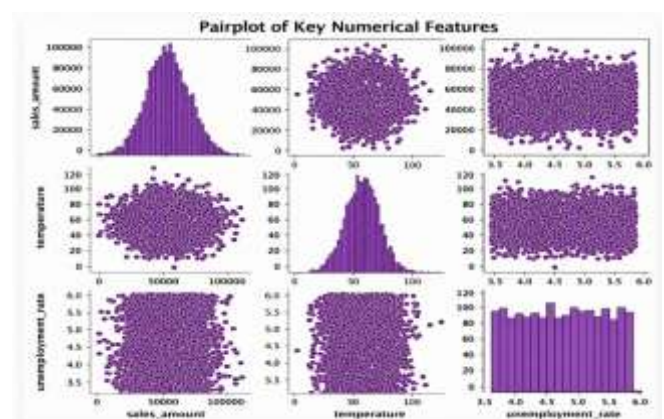


Fig. 4. Pairplot of Key Numerical Features

Fig. 4 presents the pairwise relationships and distributions of sales\_amount, temperature, and unemployment\_rate. The diagonal plots show the individual distributions, while the scatter plots show relationships between each pair of numerical features. The plots indicate the overall spread and possible relationships among the selected variables.

## B. Data Pre-Processing

The Walmart Store Sales Forecasting dataset was preprocessed by handling missing values, removing duplicates, and managing outliers. Categorical features were one-hot encoded, numerical features were normalized using Min-Max scaling, and SelectKBest was used to select important features. The main preprocessing procedures are summarized as follows:

- **Missing Data Treatment:** Missing or null values were examined to maintain the consistency and completeness of the dataset. The missing values in the Markdown columns were replaced with **0**, representing the absence of markdown activity.
- **Data Redundancy Elimination:** The dataset was examined for repeated records to prevent redundancy and maintain data integrity. Duplicate observations were removed when identified to ensure that each record contributed independently to the analysis.
- **Extreme Value Management:** The dataset was analyzed to detect unusually high or low observations that could influence the learning process. Extreme values were removed or appropriately managed based on the characteristics of the retail data.

## C. One-Hot Encoding (Binary Encoding)

One-Hot Encoding (Binary Encoding) is a preprocessing technique that converts categorical variables into separate binary features. Each category is represented by 1 when present and 0 when absent, making categorical data suitable for machine learning models.

## D. Min-Max Normalization

The dataset was normalized using Min-Max normalization, which scales feature values to the range [0, 1]. This preprocessing step helps improve model performance by normalizing features to a common scale and reducing the influence of extreme values. The normalization process was performed using the following mathematical formula Equation (1):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \square \square \square$$

where  $X$  denotes the original feature value,  $X'$  represents the normalized feature value,  $X_{min}$  indicates the minimum value of the feature, and  $X_{max}$  denotes the maximum value of the corresponding feature.

## E. Feature Selection Using SelectKBest Method

Feature selection identifies the most significant variables for prediction by removing irrelevant or less informative features. Using SelectKBest, features are ranked based on their statistical scores with respect to the target variable. Fig. 5 shows that Store has the highest score, followed by Temperature, Unemployment, CPI, and is\_thanksgiving. Fuel\_Price, is\_christmas, is\_super\_bowl, and is\_labour\_day have minimal or zero scores, indicating limited contribution to prediction.

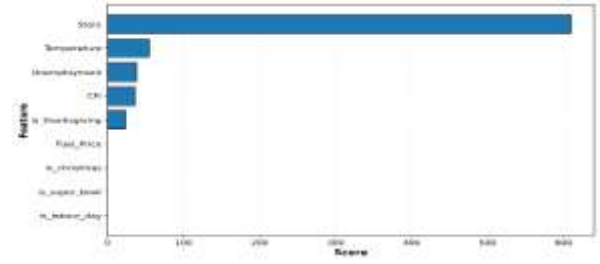


Fig. 5. Feature Importance in Walmart Sales Prediction

## F. Data Splitting

In this study, the dataset was divided into training and testing sets using an 80:20 split, where 80% of the data was used for model training and the remaining 20% was used for testing.

## G. Proposed Random Forest (RF) Model

A Random Forest (RF) model is proposed for predictive price analytics in intelligent retail management. RF is an ensemble learning algorithm that combines multiple decision trees to improve regression performance. Each tree is trained using random subsets of data and features, increasing model diversity. For regression, the final prediction is obtained by averaging the outputs of all trees. The data are recursively divided based on feature conditions, with Gini impurity used to evaluate the quality of each split. The Gini impurity at node  $N$  is given by Equation (2).

$$g(N) = \sum_{i \neq j} P(w_i)P(w_j) \quad (2)$$

where  $P(w_i)$  is the proportion of the population with class label  $i$ . Another function that can be used to judge the quality of split is Shannon Entropy. It measures the disorder in the information content. In Decision trees, Shannon entropy is used to measure the unpredictability in the information contained in a particular node of a tree (In this context, it measures how mixed the population in a node is). The entropy in a node  $N$  can be calculated as follows Equation (3):

$$H(N) = \sum_{i=1}^d P(w_i) \log_2(Pw_i) \quad (3)$$

where  $d$  is number of classes considered and  $P(w_i)$  is the proportion of the population labeled as  $i$ . Entropy is the highest when all the classes are contained in equal proportion in the node. It is lowest when there is only one class present in a node (when the node is pure). The obvious heuristic for choosing the best splitting decision at a node is the one that reduces impurity as much as possible. In other words, the best split is characterized by the highest gain in information or the highest reduction in impurity. The information gain due to a split can be calculated as follows Equation (4):

$$\Delta I(N) = I(N) - P_L * I(N_L) - P_R * I(N_R) \quad (4)$$

where  $I(N)$  is the impurity measure (Gini or Shannon Entropy) of node  $N$ ,  $P_L$  is the proportion of the population in node  $N$  that goes to the left child of  $N$  after the split and similarly,  $P_R$  is the proportion of the population in node  $N$  that goes to the right child after the split.  $N_L$  and  $N_R$  are the left and right children of  $N$  respectively. The proposed Random Forest (RF) model was configured with `n_estimators = 100`, `max_depth = None`, and `random_state = 42` to ensure stable and reproducible

predictions. The model used `min_samples_split = 2`, `min_samples_leaf = 1`, and `max_features = 1.0` for tree construction and feature selection at each split. The criterion = `squared_error` was employed as the regression loss function, with `bootstrap = True` for ensemble learning.

#### H. Evaluation Metrics

Model evaluation is an essential stage in machine learning for assessing predictive performance. In this study, four regression metrics, namely R-square ( $R^2$ ), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), are used.  $R^2$  measures the proportion of variability explained by the model, with higher values indicating better performance. RMSE represents the square root of MSE and indicates prediction deviations. MSE measures the average squared difference between actual and predicted values, while MAE measures their average absolute difference. Lower MSE, RMSE, and MAE values indicate fewer prediction errors. The equations for these metrics are shown in Equations (5)–(8):

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_i^p)^2} \quad (6)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y_i^p)^2 \quad (7)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y_i^p| \quad (8)$$

Where:

$n$  is the total number of observations.

$y_i$  is the actual Real Estate Prices at time  $i$ .

$y_i^p$  is the predicted Real Estate Prices at time  $i$ .

## IV. RESULTS AND DISCUSSION

This section describes the experimental setup and highlights the performance of the proposed model across the evaluation process, demonstrating its effectiveness and efficient processing. The research was conducted in Google Colab, utilizing 12.7 GB RAM and a Xeon CPU for efficient computation. Python 3.10 was used along with Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn for data preprocessing, model development, evaluation, and visualization. The proposed Random Forest (RF) model was trained using the Walmart Store Sales Forecasting Dataset to predict retail prices effectively. The model performance was evaluated using  $R^2$ , RMSE, MAE, and MSE as the primary regression metrics. As presented in Table II, the proposed RF model achieved an  $R^2$  value of 98.96%, indicating a strong fit between predicted and actual values. The model further obtained an RMSE of 0.488, MAE of 0.317, and MSE of 2.386, demonstrating accurate and reliable predictive performance for intelligent retail price analytics.

TABLE II. PERFORMANCE EVALUATION OF THE PROPOSED MODEL FOR PREDICTIVE PRICE ANALYTICS IN INTELLIGENT RETAIL MANAGEMENT

| Matrix    | Random Forest (RF) |
|-----------|--------------------|
| R-Squared | 98.96              |
| RMSE      | 0.488              |
| MAE       | 0.317              |
| MSE       | 2.386              |

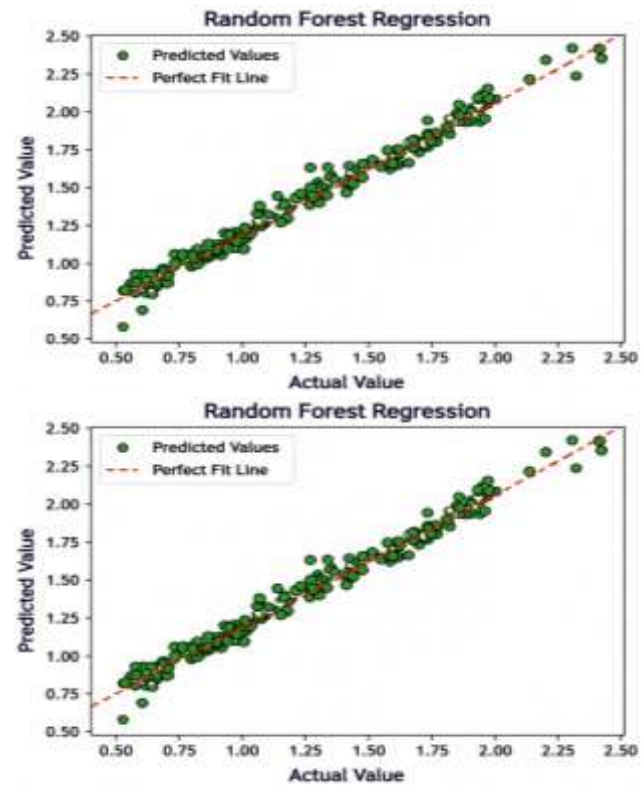


Fig. 6. Scatter plot for Actual vs. Predicted Values for Random Forest Regression

Fig. 6 shows the relationship between the actual and predicted values of the Random Forest regression model. Most points are close to the perfect fit line, indicating that the predicted values are close to the actual values. Overall, the model demonstrates good prediction performance with relatively small errors.

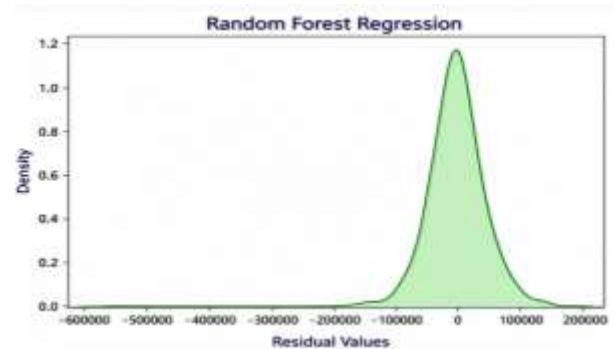


Fig. 7. Residual Density Plot for Random Forest Regression

Fig. 7 shows the distribution of residual values produced by the Random Forest Regression model. Most residuals are

concentrated around zero, indicating that the predicted values are generally close to the actual values. The relatively narrow distribution suggests low prediction error and good model performance.

#### A. Comparative Analysis

To assess the effectiveness of the proposed model, a comparative accuracy evaluation against other existing models is provided in Table III. The results compare OLS, XGBoost, Logistic Regression (LR), and the proposed Random Forest (RF) model using R<sup>2</sup>, RMSE, MAE, and MSE. OLS achieved an R<sup>2</sup> of 72.57%, while XGBoost and LR achieved 97.52% and 97.60%, respectively. In comparison, the proposed RF model obtained the highest R<sup>2</sup> value of 98.96%, demonstrating superior explanatory capability. The proposed model also achieved an RMSE of 0.488, MAE of 0.317, and MSE of 2.386, indicating effective predictive performance for retail price prediction. Overall, the comparative results demonstrate that the proposed RF model provides a strong and competitive approach for intelligent retail price analytics.

TABLE III. COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR RETAIL PRICE PREDICTION

| Model       | R-Square d | RMSE      | MAE       | MSE      |
|-------------|------------|-----------|-----------|----------|
| OLS[25]     | 72.57      | -         | 0.6728    | 1.1135   |
| XGBoost[26] | 97.52      | 0.09394   | 0.06188   | 0.008826 |
| LR[27]      | 97.60      | 80153.858 | 35632.510 | -        |
| Proposed RF | 98.96      | 0.488     | 0.317     | 2.386    |

The proposed Random Forest (RF) model achieves the highest R-squared value of 98.96%, demonstrating superior predictive accuracy compared with the existing models. It has the ability to capture complex and nonlinear relationships within the retail price data, due to its ensemble learning approach. The model is stable and reliable and can be used for a variety of data types and product properties. Its highly predictive nature can aid accurate pricing and better retail analytics. The model can help retailers with more effective pricing and stock management decisions as well. Overall, the proposed RF model offers an accurate, robust, and effective solution for intelligent retail price prediction.

## V. CONCLUSION AND FUTURE STUDY

The prediction of product prices in the retail industry is a major problem, and its application would be useful in the business sector for making effective pricing decisions in a competitive environment. This paper focuses on a new approach for predicting product prices using regression-based machine learning techniques. The proposed Random Forest (RF) model was developed using the Walmart Store Sales Forecasting

Dataset. The performance of the proposed model was evaluated using the R-squared metric and compared with existing regression approaches. The proposed RF achieved an R<sup>2</sup> value of 98.96%, outperforming OLS (72.57%), XGBoost (97.52%), and Logistic Regression (97.60%). These results demonstrate that the proposed RF model provides highly accurate predictive performance and effectively captures important patterns within retail sales data. Therefore, the proposed approach can support intelligent retail management by enabling reliable price prediction, improving pricing strategies, and assisting businesses in making data-driven decisions. Future work can focus on incorporating larger and more diverse retail datasets, advanced optimization techniques, and real-time market information to further improve prediction accuracy and practical applicability.

## REFERENCES

1. S. Kukar, “A Generative AI-Based Framework for Retail Sales Forecasting and Demand Prediction,” in *2026 2nd International Conference on Artificial Intelligence Systems (AIS)*, IEEE, May 2026, pp. 306–311. doi: 10.1109/AIS69919.2026.11621469.
2. M. Spuritha, C. S. Kashyap, T. R. Nambiar, D. R. Kiran, N. S. Rao, and G. P. Reddy, “Quotidian Sales Forecasting using Machine Learning,” in *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, IEEE, Sep. 2021, pp. 1–6. doi: 10.1109/ICSES52305.2021.9633975.
3. N. Bhatt, “A Hybrid Artificial Intelligence Model for Product Recommendation in Online Retail System,” in *2026 2nd International Conference on Artificial Intelligence Systems (AIS)*, San Antonio, TX, USA: IEEE, 2026, pp. 33–39, May. doi: 10.1109/AIS69919.2026.11621290.
4. S. Singh, S. A. Pahune, P. Chatterjee, and R. Sura, “Advanced Machine Learning Techniques for Prediction of Customer Churn in Telecommunication Sector,” in *2025 IEEE 6th India Council International Subsections Conference (INDISCON)*, Rourkela, India: IEEE, 2025, pp. 1–6, August. doi: 10.1109/INDISCON66021.2025.11252233.
5. K. Singhal, V. Singh, and A. Kaul, “Smart Retail: Utilizing Machine Learning for Demand Prediction, Price Strategy, and Inventory Management,” in *2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, Dec. 2024, pp. 485–489. doi: 10.1109/CICN63059.2024.10847534.
6. K. Muniyanayaka, “Machine Learning-based Pricing Optimization for Dynamic Pricing in Online Retail,” *J. Informatics Educ. Res.*, vol. 4, no. 3, pp. 2145–2151, 2024, doi: 10.52783/jier.v4i3.1564.
7. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and Machine Learning forecasting

- methods: Concerns and ways forward,” *PLoS One*, vol. 13, no. 3, p. e0194889, Mar. 2018, doi: 10.1371/journal.pone.0194889.
8. P. Liu, M. Li, D. Dai, and L. Guo, “The effects of social commerce environmental characteristics on customers’ purchase intentions: The chain mediating effect of customer-to-customer interaction and customer-perceived value,” *Electron. Commer. Res. Appl.*, vol. 48, p. 101073, Jul. 2021, doi: 10.1016/j.elerap.2021.101073.
9. N. Ahmed Mahoto, R. Iftikhar, A. Shaikh, Y. Asiri, A. Alghamdi, and K. Rajab, “An Intelligent Business Model for Product Price Prediction Using Machine Learning Approach,” *Intell. Autom. Soft Comput.*, vol. 29, no. 3, pp. 147–159, 2021, doi: 10.32604/iasc.2021.018944.
10. D. Patel, “Integrating Price Elasticity and Reinforcement Learning: A Data-Driven Framework for Strategic E-commerce Pricing,” in *2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC)*, Houston, TX, USA: IEEE, 2026, pp. 1–6, February. doi: 10.1109/ICAIC67076.2026.11395747.
11. S. Kumar, S. Rani, A. Jain, M. Kumar, and P. Jaglan, “Automatic Face Mask Detection Using Deep Learning-Based Mobile-Net Architecture,” in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, Sep. 2023, pp. 1075–1080. doi: 10.1109/IC3I59117.2023.10397772.
12. N. B. Gaikwad *et al.*, “Hardware Design and Implementation of Multiagent MLP Regression for the Estimation of Gunshot Direction on IoBT Edge Gateway,” *IEEE Sens. J.*, vol. 23, no. 13, pp. 14549–14557, Jul. 2023, doi: 10.1109/JSEN.2023.3278748.
13. W. DeGroat, D. Mendhe, A. Bhusari, H. Abdelhalim, S. Zeeshan, and Z. Ahmed, “IntelliGenes : a novel machine learning pipeline for biomarker discovery and predictive analysis using multi-genomic profiles,” *Bioinformatics*, vol. 39, no. 12, Dec. 2023, doi: 10.1093/bioinformatics/btad755.
14. F. Dakalbab, M. A. Talib, Q. Nasir, and T. Saroufil, “Artificial intelligence techniques in financial trading: A systematic literature review,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 3, p. 102015, Mar. 2024, doi: 10.1016/j.jksuci.2024.102015.
15. [15] B. Ho, C. Pham-Quoc, T. N. Thinh, and N. Thoai, “A Secured OpenFlow-Based Switch Architecture,” in *2016 International Conference on Advanced Computing and Applications (ACOMP)*, IEEE, Nov. 2016, pp. 83–89. doi: 10.1109/ACOMP.2016.021.
16. S. Dixit and N. Soni, “Enhancing stock market prediction using three-phase classifier and EM-EPO optimization with news feeds and historical data,” *Multimed. Tools Appl.*, vol. 83, no. 13, pp. 37859–37887, Oct. 2023, doi: 10.1007/s11042-023-17184-x.
17. S. Kukkar, “Leveraging Price-Quote Data with AI for Smarter Configure-Price-Quote Systems,” in *2026 2nd International Conference on Artificial Intelligence Systems (AIS)*, San Antonio, TX, USA: IEEE, 2026, pp. 329–335, May. doi: 10.1109/AIS69919.2026.11621526.
18. [G. Preethika, T. Keshav, C. S. Yadav, Y. Sunanda, G. B. Regulwar, and P. N. Reddy, “Dynamic Pricing for Retail Products Using AI,” in *2026 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, IEEE, Jan. 2026, pp. 1–7. doi: 10.1109/IITCEE67948.2026.11394359.
19. I. Admeur and R. Zine, “Hybrid SARIMAX–SVR Forecasting of Retail Diesel Prices in Morocco,” in *2026 6th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, IEEE, May 2026, pp. 1–5. doi: 10.1109/IRASET68627.2026.11538718.
20. F. R. Padila, S. Yuliyanti, and M. Al Husaini, “A Hybrid Content-Based and Stacking Ensemble Framework for Forecasting and Recommending Product Price Trends,” in *2025 Tenth International Conference on Informatics and Computing (ICIC)*, IEEE, Oct. 2025, pp. 1–6. doi: 10.1109/ICIC68054.2025.11309470.
21. K. Poongothai, D. G. S. B. D, and Y. S., “Smart Retail Using Machine Learning for Demand Forecasting and Inventory Optimization,” in *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICES)*, IEEE, Dec. 2024, pp. 1–7. doi: 10.1109/ICES63760.2024.10910874.
22. K. Chen, X. Liang, Y. Mo, Y. Li, L. Xu, and W. Kong, “A Retail Product Prediction Model Incorporating Machine Learning and Propensity Score Matching Method,” in *2024 7th International Conference on Computer Information Science and Application Technology (CISAT)*, IEEE, Jul. 2024, pp. 1030–1038. doi: 10.1109/CISAT62382.2024.10695432.
23. H. Upadhyay, S. Shekhar, A. Vidyarthi, R. Prakash, and R. Gowri, “Sales Prediction in the Retail Industry Using Machine Learning: A Case Study of BigMart,” in *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, IEEE, Aug. 2023, pp. 1–6. doi: 10.1109/ELEXCOM58812.2023.10370313.
24. C. McRae, J.-Y. Nie, and L. Dube, “Heterogeneity in feature importance and prediction performance for sales at the market and store levels: the case of branded yogurt products in Quebec,” in *2022 IEEE International Conference on Big Data (Big Data)*,

- IEEE, Dec. 2022, pp. 4088–4096.  
doi: 10.1109/BigData55660.2022.10020567.
25. T. Yang, “Sales Prediction of Walmart Sales Based on OLS, Random Forest, and XGBoost Models,” *Highlights Sci. Eng. Technol.*, vol. 49, pp. 244–249, May 2023, doi: 10.54097/hset.v49i.8513.
26. P. Bansal, “Role of Artificial Intelligence (AI)-Driven Demand Forecasting: A Machine Learning Approach for Supply Chain Resilience,” *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 4, pp. 4056–4067, May 2025, doi: 10.38124/ijisrt/25apr2260.
27. C. N. C. *et al.*, “Advancing Retail Predictions: Integrating Diverse Machine Learning Models for Accurate Walmart Sales Forecasting,” *Asian J. Probab. Stat.*, vol. 26, no. 7, pp. 1–23, Jun. 2024, doi: 10.9734/ajpas/2024/v26i7626.