IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2026

DOI: 10.17148/IJIREEICE.2026.14711

Machine Learning Based Retail Profit Prediction for Business Decision Support – A Comparative Model Analysis

Purnithaa B R[1], Maneeswar K G[2], Dr. Roselin A[3]

III B.Sc. Computer Science with Data Analytics,

Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India[1,2]

Assistant Professor, Department of Computer Science with Data Analytics,

Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India[3]

Abstract: Retail organizations generate large volumes of transactional data, making accurate profit estimation essential for effective business decision-making. This study proposes a machine learning-based framework for predicting order-level retail profit using the SuperStore Orders dataset containing 51,290 records and 21 attributes. Three regression models—Linear Regression, Decision Tree Regressor, and Random Forest Regressor—were developed and evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). Experimental results demonstrate that the Random Forest Regressor achieved the highest predictive performance with an R^2 score of 0.7095, outperforming the other models by effectively capturing complex non-linear relationships among retail transaction features. The findings indicate that Random Forest is a reliable approach for retail profit prediction and can serve as a valuable predictive component in Business Decision Support Systems.

Keywords: Machine Learning, Random Forest, Retail Profit Prediction, Business Decision Support System, Regression Analysis

I. INTRODUCTION

The retail sector operates in an environment characterized by high transaction volume, thin margins, and constantly shifting consumer demand, making profitability management one of the most persistent challenges faced by retail organizations. Every sale carries a combination of factors—selling price, quantity, discount, shipping cost, and the category or region of the transaction—that jointly determine whether it contributes a profit or a loss. Traditional decision-making in retail has historically relied on descriptive reporting and managerial intuition, an approach that becomes increasingly inadequate as the scale and complexity of transactional data grows. Conventional Business Decision Support Systems (DSS), built on static rules and descriptive dashboards, struggle to anticipate outcomes such as profitability, which is jointly influenced by non-linear, interacting factors that are difficult to express through simple formulae.

Machine Learning (ML) offers a natural extension to DSS by learning these complex relationships directly from data. This study addresses the problem of predicting order-level profit in a large, multi-feature retail transactions dataset, comparing multiple regression-based ML models to identify which is best suited as a predictive component of a retail Business Decision Support System. The SuperStore Orders dataset sourced from Kaggle [10] (51,290 records, 21 attributes) is used as the empirical basis of this work; three regression models—Linear Regression, Decision Tree Regressor, and Random Forest Regressor—are trained on fourteen selected input features and evaluated using MAE, RMSE, and R^2. Prediction is framed at the point an order is finalized—when its sales value, quantity, discount, and shipping cost are already known but Profit has not yet been formally calculated—a distinction discussed further in Section III. Section II reviews related literature and identifies the research gap; Section III describes the dataset and methodology; Section IV presents and discusses the comparative results; and Section V concludes with implications for practical deployment.

II. LITERATURE SURVEY

Machine learning has increasingly been applied across business analytics, retail forecasting, and decision support system research, motivated by the growing availability of transactional data and the demonstrated ability of ensemble methods to model non-linear business relationships. Table 2 summarizes five representative studies published between 2023 and 2025 addressing business decision support, retail profit and sales prediction, machine learning in business analytics, and Random Forest-based prediction.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2026

DOI: 10.17148/IJIREEICE.2026.14711

TABLE II LITERATURE SURVEY COMPARISON

Paper Title	Authors	Year	Method	Key Findings	Research Gap
Retail Market Firm Performance Prediction	Vukovic et al. [1]	2023	Regression, DNN, LSTM, RF, Ensembles	RF/ensembles outperform regression for profitability (ROA)	Firm-level, not order-level data
Retail Demand Prediction: Tree Ensembles vs LSTM	Nasseri et al. [2]	2023	ETR, RF, GBR, XGBoost, LSTM	Tree ensembles outperform LSTM in MAE, RMSE, R^2	Demand, not profit
Big Mart Sales Prediction (LR, RF, GB)	Kang [3]	2023	LR, RF, Gradient Boosting	LR underfits; tree models perform better	Small dataset; sales only
Business Analytics & Decision Science Review	Ibeh et al. [4]	2024	Review of BA/DS techniques	ML/AI central to strategic decisions	Conceptual, no benchmark
Retail Sales Forecasting with ML	Mustapha & Sithole [5]	2025	RF, XGBoost, LR, ARIMA, LSTM	RF/boosting outperform LR & ARIMA	Sales volume, limited scope

The reviewed literature demonstrates a consistent pattern: across firm-level profitability forecasting [1], demand forecasting [2], small-scale sales prediction [3], and general sales forecasting [5], tree-based ensembles and Random Forest in particular repeatedly outperform simple linear regression when predictors and the financial outcome are non-linearly related. Vukovic et al. [1] found ML methods, including Random Forest, produced lower forecast error than random-effects regression, while Nasseri et al. [2] showed tree-based ensembles outperformed LSTM-based deep learning for demand prediction. Kang [3] similarly found linear regression underfit a small retail sales dataset. Ibeh et al. [4] situate these findings within the broader trajectory of business analytics.

Despite this body of work, a gap remains. Profitability studies such as [1] operate at the firm level rather than individual transactions. Demand and sales-forecasting studies such as [2], [3], and [5] target sales volume rather than profit—a distinct, arguably more decision-relevant outcome, since a high-sales, high-discount order can still generate a loss. Review-oriented work such as [4] provides conceptual grounding but no quantitative benchmark. This study aims to help address that gap by predicting profit directly from a large (51,290-record), order-level, multi-feature retail dataset, benchmarking Linear Regression, Decision Tree Regressor, and Random Forest Regressor within an explicit Business Decision Support System framing.

III. METHODOLOGY

A. Dataset

This study uses the SuperStore Orders dataset obtained from Kaggle [10], a widely used, publicly available retail transactions dataset recording individual customer orders placed with a multi-category superstore operating across several international markets. The dataset consists of 51,290 records described by 21 attributes spanning order, geographic, product, and financial information. Profit, the continuous financial outcome recorded for every order, is treated as the target variable, since it reflects the net contribution of each order rather than gross revenue alone. Table 1 summarizes the dataset.

TABLE I DATASET DESCRIPTION

Attribute	Description
Dataset Name	SuperStore Orders Dataset
Source	Kaggle
Total Records	51,290
Total Features	21
Target Variable	Profit
Input Features Used	14 (Sales, Quantity, Discount, Shipping Cost, Year, Ship Mode, Segment, State, Country, Market, Region, Category, Sub-Category, Order Priority)
Data Type	Structured / Tabular
Train-Test Split Ratio	80 : 20

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2026

DOI: 10.17148/IJIREEICE.2026.14711

B. Prediction Scenario and Feature Availability

The dataset includes four fields closely tied to Profit: Sales, Quantity, Discount, and Shipping Cost. Before using these as predictors, it is necessary to state at what point in the order lifecycle prediction occurs, since this determines which fields are legitimately available. Three framings are possible: pre-order prediction (before the order exists, using only historical, customer-level, or product-level data); order-time profit estimation (the order's sales amount, quantity, discount, and shipping cost are known or decided, but Profit has not yet been calculated or posted); and post-transaction profit analysis (Profit is already recorded, so the model describes rather than predicts historical outcomes).

This study adopts the order-time profit estimation framing: the four fields are order-defining values known before Profit is computed and posted, so none constitutes target leakage in the strict sense of encoding information unavailable at prediction time. Sales warrants a separate caveat, however: Profit here is derived largely as Sales minus underlying product cost, so Sales has a strong definitional relationship with Profit that is distinct from an independent driver such as Discount or Region. Part of the reported predictive power may therefore reflect this accounting relationship rather than a fully independent pattern—a property of the chosen scenario, not a flaw to correct by removing Sales, since a retailer would reasonably know this value at order time. This is revisited in Section IV.C (Limitations).

C. Data Preprocessing and Categorical Encoding

Raw transactional data of this scale contains formatting inconsistencies: the Sales column, originally text due to thousand-separator commas, was cleaned and converted to numeric type, and the dataset was checked for missing values (none critical were found). The nine categorical attributes—Ship Mode, Segment, State, Country, Market, Region, Category, Sub-Category, and Order Priority—were transformed using Label Encoding. Following cleaning and encoding, the dataset was split 80:20 into training and test sets.

With the exception of Order Priority, which has a natural order (Low–Critical), the remaining eight attributes are nominal, and Label Encoding imposes an arbitrary numeric ordering on them that can distort Linear Regression's coefficients; tree-based models are considerably less sensitive to this, since they use encoded values only as split thresholds. One-Hot Encoding is recommended for low-cardinality attributes (e.g., Ship Mode, Segment, Category), since it avoids implying a false ordinal relationship without materially increasing dimensionality; for higher-cardinality attributes (e.g., Country, State, possibly Sub-Category or Market), One-Hot Encoding would substantially increase dimensionality, so grouping rare categories or frequency encoding may be preferable, depending on the actual cardinality present. Adopting an alternative encoding would require retraining and re-evaluating all three models on the resulting feature space; this is left as a direction for future work (Section V) rather than altering the results reported in Table 3, which reflect the original Label-Encoded experiment used throughout this study.

D. Feature Selection

From the 21 available attributes, fourteen were retained as input features on the basis of their relevance to profit: Sales, Quantity, Discount, Shipping Cost, Year, Ship Mode, Segment, State, Country, Market, Region, Category, Sub-Category, and Order Priority. Identifier columns such as Order ID, Customer Name, Product Name, and Row ID were excluded, as they carry no generalizable predictive signal.

E. Regression Models and Experimental Setup

Three regression models of increasing structural complexity were trained and compared.

1) Linear Regression: used as a baseline; it assumes an additive linear relationship between features and profit, minimizing the sum of squared residuals. It is interpretable but cannot capture non-linear feature interactions common in retail transaction data.

2) Decision Tree Regressor: recursively partitions the feature space by selecting, at each node, the feature and threshold that most reduce prediction error [7]. It models non-linear relationships but is prone to overfitting when grown to full depth.

3) Random Forest Regressor: an ensemble of many decision trees, each trained on a bootstrap-resampled subset of the data and a random subset of features, whose predictions are averaged, building on established random forest and random decision forest methods [6], [11]. This reduces the variance of any single tree and generally improves generalization on data combining numeric and categorical features with non-linear relationships.

All experiments were implemented in Python using pandas and NumPy for data handling and scikit-learn for modelling [8]. The 80:20 train-test split was produced by scikit-learn's train_test_split with test_size = 0.2 and random_state = 42. Linear Regression used scikit-learn's default parameters (no tuning). Decision Tree Regressor used random_state = 42 with all other parameters at their scikit-learn defaults. Random Forest Regressor used n_estimators = 100 and random_state = 42, with all other parameters at their scikit-learn defaults. Reproduction of the reported results (Table 3) was independently verified using scikit-learn 1.8.0, pandas 3.0.2, and NumPy 2.4.4, which returned MAE, RMSE, and R^2 values identical to those reported for all three models; the exact library versions used in the original experiment were

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2026

DOI: 10.17148/IJIREEICE.2026.14711

not recorded in the source code and remain unconfirmed, though the exact match obtained here indicates the results are robust to this.

F. Performance Evaluation and Cross-Validation

The trained models were evaluated on the held-out test set using three complementary metrics [9], [14], where y_i is the actual profit, $\hat{y}_i$ the predicted profit, $\bar{y}$ the mean actual profit, and n the number of test observations.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Lower MAE and RMSE values, together with an R^2 value closer to 1, indicate stronger predictive performance.

Table 3 is based on a single 80:20 split, which provides only one sample of performance and can be sensitive to which records fall into the test set, especially given Profit's skewed distribution (Fig. 2) and outliers. To assess this, 5-fold cross-validation was additionally performed on the same feature set and Label Encoding as the original experiment, using scikit-learn's KFold with 5 splits, shuffle = True, and random_state = 42: the training data was partitioned into five folds, each model was trained on four folds and evaluated on the fifth, this was repeated five times so each fold served once as the validation set, and MAE, RMSE, and R^2 were averaged across folds along with their standard deviation. Results are reported in Table 4 and discussed in Section IV.B.

IV. RESULTS & DISCUSSION

A. Train-Test Split Results

The three regression models were trained on the same 80 percent training partition and evaluated on the identical 20 percent test partition, ensuring a fair, like-for-like comparison. Table 3 summarizes the MAE, RMSE, and R^2 achieved by each model.

TABLE III PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	MAE	RMSE	R^2
Linear Regression	59.85	141.79	0.3874
Decision Tree Regressor	45.93	131.53	0.4728
Random Forest Regressor	**35.94**	**97.63**	**0.7095**

Linear Regression produced the weakest performance (MAE = 59.85, RMSE = 141.79, R^2 = 0.3874), explaining only around 39 percent of the variance in order-level profit, consistent with its assumption that profit depends additively and linearly on the input features. The Decision Tree Regressor improved substantially (R^2 = 0.4728) by modelling non-linear splits and feature interactions. The Random Forest Regressor clearly outperformed both, achieving the lowest MAE (35.94) and RMSE (97.63) and the highest R^2 (0.7095), explaining approximately 71 percent of the variance in profit on unseen test data.

This improvement stems from Random Forest's ensemble structure: averaging predictions from many bootstrapped trees reduces the variance and overfitting of a single Decision Tree while retaining the ability to model non-linear, interaction-heavy relationships between Sales, Discount, Shipping Cost, and categorical attributes such as Region and Category that are jointly associated with Profit [6]—predictive associations rather than evidence of causal relationships. As Fig. 4 shows, Random Forest's predictions track actual profit closely for orders clustered near zero, but compress toward zero for the rarest, most extreme profit and loss observations—a common limitation of tree ensembles on skewed targets like Fig. 2's, discussed further in Section IV.C. Random Forest therefore represents the most suitable of the three models to serve as a predictive component within a retail Business Decision Support System, offering a substantially more reliable profit estimate than the simpler alternatives, while remaining less reliable for unusually large gains or losses.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2026

DOI: 10.17148/IJIREEICE.2026.14711

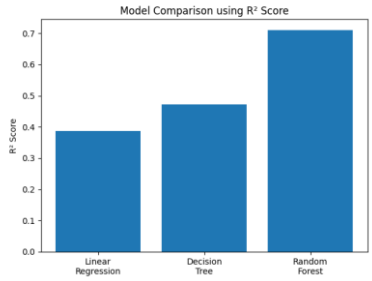


Fig. 1. Model Comparison using R^2 Score

Random Forest reaches an R^2 of ≈ 0.71, clearly above Decision Tree (≈ 0.47) and Linear Regression (≈ 0.39), visually confirming its superior explanatory power.

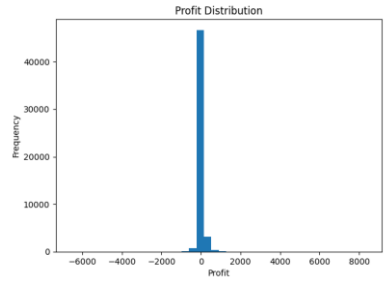


Fig. 2. Profit Distribution

Profit is sharply concentrated near zero (over 46,000 orders), with a much smaller spread into losses (down to ≈ −700) or gains (up to ≈ 1,000) and rare extreme outliers beyond that range.

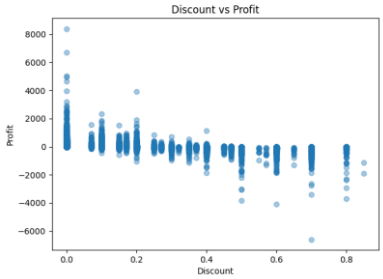


Fig. 3. Discount vs Profit

Low-discount orders show wide, mostly positive profit (up to ≈ 8,400); beyond discounts of ≈ 0.4–0.5 the cloud shifts almost entirely to losses, reaching ≈ −6,600 at 0.6–0.8 discount.

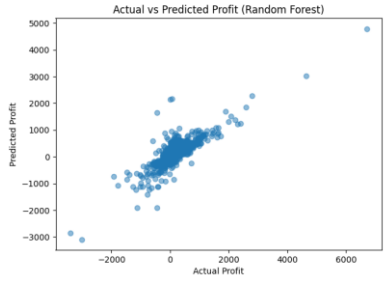


Fig. 4. Actual vs Predicted Profit (RF)

Points cluster tightly along the diagonal near the origin; high-profit and high-loss outliers are predicted in the correct direction but with somewhat compressed magnitude, consistent with R^2 = 0.7095.

B. Five-Fold Cross-Validation

Table 4 reports the mean and standard deviation of MAE, RMSE, and R^2 for each model across five folds of cross-validation, using the same feature set, Label Encoding, and random_state = 42 as the original experiment. The relative ranking of the three models under cross-validation matches the single 80:20 split exactly: Random Forest Regressor achieves the lowest mean MAE and RMSE and the highest mean R^2 (0.7031), followed by Decision Tree Regressor (R^2 = 0.4549), with Linear Regression weakest (R^2 = 0.3059). This confirms that the original train-test split's conclusion—that Random Forest is the best-performing model—is not an artifact of a single favorable split.

TABLE IV FIVE-FOLD CROSS-VALIDATION PERFORMANCE

Model	MAE (Mean ± SD)	RMSE (Mean ± SD)	R^2 (Mean ± SD)
Linear Regression	58.82 ± 0.64	144.77 ± 10.19	0.3059 ± 0.0782
Decision Tree Regressor	45.13 ± 1.55	127.25 ± 6.55	0.4549 ± 0.1206
Random Forest Regressor	34.39 ± 1.03	94.62 ± 3.99	0.7031 ± 0.0286

Random Forest also shows the smallest relative variability across folds (R^2 standard deviation = 0.0286), indicating it generalizes most consistently across different partitions of the data, whereas Decision Tree Regressor shows the largest variability (SD = 0.1206), consistent with a single tree's known sensitivity to the specific training sample. For Random Forest, the cross-validated mean R^2 (0.7031) is close to the original single-split value (0.7095), supporting the reliability of the reported result. For Linear Regression, however, the cross-validated mean R^2 (0.3059) is noticeably lower than the original single-split value (0.3874), indicating that the original 80:20 split happened to be somewhat more favorable to Linear Regression than is typical; its single-split R^2 should therefore be read as a relatively optimistic estimate rather than a fully representative one. This does not change the paper's central conclusion, since Random Forest remains clearly the best-performing model under both evaluation schemes, but it illustrates precisely why cross-validation is a useful complement to a single train-test split.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 7, July 2026

DOI: 10.17148/IJIREEICE.2026.14711

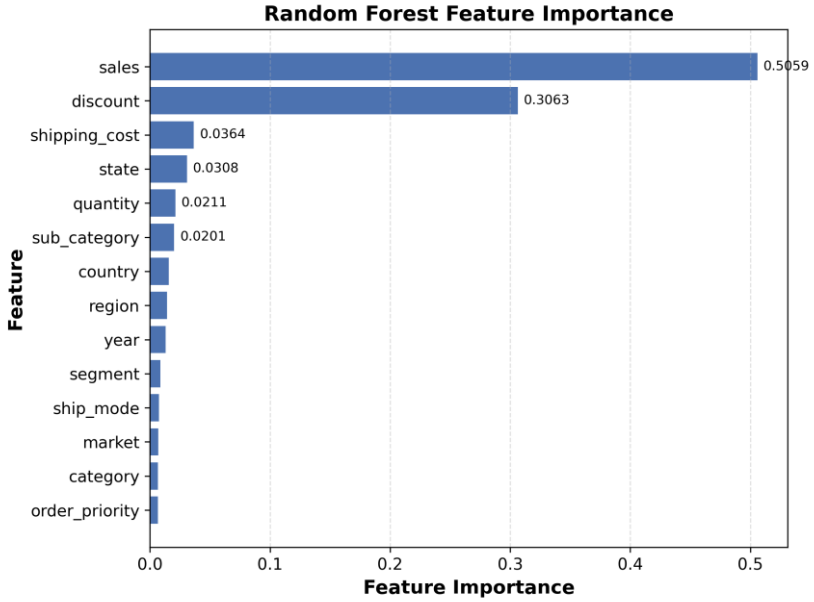
C. Random Forest Feature Importance

Table 5 and Fig. 5 present the ranked feature-importance scores extracted from the trained Random Forest Regressor's feature_importances_ attribute [8], which sum to 1.0000 across the fourteen input features. Sales was most strongly associated with model predictions (importance = 0.5059), followed by Discount (0.3063); together these two features account for over 81% of the total importance captured by the model. Shipping Cost (0.0364) and State (0.0308) were the next most influential, followed by Quantity (0.0211) and Sub-Category (0.0201). The remaining geographic, temporal, and categorical features—Country, Region, Year, Segment, Ship Mode, Market, Category, and Order Priority—each individually contributed less than 2% of total importance.

TABLE V RANDOM FOREST FEATURE IMPORTANCE RANKING

Rank	Feature	Importance
1	Sales	0.5059
2	Discount	0.3063
3	Shipping Cost	0.0364
4	State	0.0308
5	Quantity	0.0211
6	Sub-Category	0.0201
7	Country	0.0158
8	Region	0.0143
9	Year	0.0131
10	Segment	0.0088
11	Ship Mode	0.0074
12	Market	0.0070
13	Category	0.0065
14	Order Priority	0.0065

Fig. 5. Random Forest Feature Importance



This ranking is consistent with, and provides direct empirical support for, the discussion in Section III.B: Sales' outsized importance reflects its close definitional relationship with Profit in this dataset, while Discount's substantial importance is consistent with the negative, non-linear relationship between discount level and profit observed in Fig. 3. From a business decision support perspective, this ranking suggests that discount-policy decisions—rather than shipping,

DOI: 10.17148/IJIREEICE.2026.14711

geographic, or product-category decisions—are the input a manager can most directly influence to affect predicted profit, since Sales itself is largely a function of the transaction rather than a discretionary lever. These findings represent predictive associations with model outputs and should not be interpreted as evidence that any individual feature causes changes in profit.

D. Limitations

This study has several limitations. It relies on a single public dataset (SuperStore Orders via Kaggle), so relative model performance may not generalize to other retailers. Only three regression algorithms were tested, excluding boosting methods such as Gradient Boosting [13] and XGBoost [12]. Label Encoding of nominal attributes (Section III.C) remains a particular concern for Linear Regression, whose cross-validated R^2 (Section IV.B) was also noticeably more variable across folds than Random Forest's. Four input features (Sales, Quantity, Discount, Shipping Cost) have a close temporal—and for Sales, definitional—relationship to Profit (Section III.B); the feature-importance results in Section IV.C provide direct evidence for this concern, since Sales and Discount together account for over 81% of Random Forest's total importance, so part of the reported accuracy reflects this proximity rather than a fully independent pattern. Profit is heavily concentrated near zero with rare extreme outliers (Fig. 2), and all three models, Random Forest included, tend to under-predict the magnitude of these extremes. Finally, this study evaluates a predictive ML component only; no complete software Business Decision Support System was implemented or externally validated.

V. CONCLUSION

This study presented a machine learning-based approach for predicting order-level retail profit using the SuperStore Orders dataset containing 51,290 records and 21 attributes. Three regression algorithms—Linear Regression, Decision Tree Regressor, and Random Forest Regressor—were developed and evaluated to identify the most effective model for retail profit prediction. Performance evaluation using MAE, RMSE, and R^2 showed that the Random Forest Regressor achieved the best results, with an MAE of 35.94, RMSE of 97.63, and an R^2 score of 0.7095. Furthermore, 5-fold cross-validation confirmed the stability and reliability of the Random Forest model, demonstrating consistent performance across different data partitions.

Feature-importance analysis revealed that Sales and Discount were the most influential variables in predicting profit, highlighting their significance in retail decision-making. The proposed model can assist organizations in estimating the profitability of customer orders, identifying potentially low-profit transactions, and supporting pricing and discount strategies. Although the study was conducted using a single public dataset and limited to three regression algorithms, the results demonstrate the effectiveness of ensemble learning for retail profit prediction. Future work may explore advanced algorithms such as Gradient Boosting and XGBoost, alternative feature encoding techniques, explainable AI methods such as SHAP [15], validation using real-world retail datasets, and the integration of the prediction model into a real-time Business Decision Support System.

REFERENCES

[1] D. B. Vukovic, L. Spitsina, E. Gribanova, V. Spitsin, and I. Lyzin, "Predicting the performance of retail market firms: Regression and machine learning methods," Mathematics, vol. 11, no. 8, art. 1916, Apr. 2023.

[2] M. Nasseri, T. Falatouri, P. Brandtner, and F. Darbanian, "Applying machine learning in retail demand prediction—A comparison of tree-based ensembles and long short-term memory-based deep learning," Applied Sciences, vol. 13, no. 19, art. 11112, Oct. 2023.

[3] R. Kang, "Sales prediction of Big Mart based on linear regression, random forest, and gradient boosting," in Proc. 2nd Int. Conf. Business and Policy Studies, vol. 17, pp. 200–207, Sep. 2023.

[4] C. V. Ibeh, O. F. Asuzu, T. Olorunsogo, O. A. Elufioye, N. L. Nduubuisi, and A. I. Daraojimba, "Business analytics and decision science: A review of techniques in strategic business decision making," World Journal of Advanced Research and Reviews, vol. 21, no. 2, pp. 1761–1769, Feb. 2024.

[5] O. O. Mustapha and T. Sithole, "Forecasting retail sales using machine learning models," American Journal of Statistics and Actuarial Sciences, vol. 6, no. 1, pp. 35–67, Apr. 2025.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Boca Raton, FL, USA: CRC Press, 1984.

[8] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

DOI: 10.17148/IJIREEICE.2026.14711

[9] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[10] L. Anwer, "Superstore Sales Dataset," Kaggle Datasets. [Online]. Available: https://www.kaggle.com/datasets/laibaanwer/superstore-sales-dataset. Accessed: Jul. 24, 2026.

[11] T. K. Ho, "Random decision forests," in Proc. 3rd Int. Conf. Document Analysis and Recognition, vol. 1, Montreal, QC, Canada, Aug. 1995, pp. 278–282.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[13] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[14] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, 2nd ed. New York, NY, USA: Springer, 2021.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 4765–4774.

