ISSN: 2349-6002

Sales Forecasting Using Machine Learning Algorithms: A Comparative Study Of Decision Tree, Random Forest, And Extra Tree

Anshul[1], Tanya Singh[2], Kushagra Sharma[3], Deepak Kumar Pathak[4]
[1,2,3,4]*Department of Computer Applications, Invertis University, Bareilly, India*
doi.org/10.64643/IJIRTV12I11-207260-459

***Abstract*—Sales forecasting is a critical function in business analytics because it supports inventory planning, procurement, budgeting, workforce allocation, and marketing decisions. In retail and food-and-beverage environments, sales are influenced by seasonality, trend, calendar effects, and short-term fluctuations, making simple forecasting rules insufficient for many practical settings.**

This study presents a comparative machine learning framework for forecasting monthly sales using three tree-based regression algorithms: Decision Tree, Random Forest, and Extra Trees.

The study follows the workflow indicated in the supplied draft, preserving its original focus on historical sales prediction through date-derived features, lag variables, data cleaning, and supervised model evaluation. The paper extends the draft into a research-style article by grounding the methodology in standard forecasting and machine learning literature. It argues that while a single decision tree is easy to interpret, ensemble models are better suited to nonlinear sales behavior because they reduce instability and improve generalization. Consistent with the direction reported in the supplied study, Extra Trees emerges as the strongest overall performer, while Random Forest remains highly competitive and robust.

The findings show that machine learning can provide practical value for sales forecasting when supported by careful preprocessing, feature engineering, and appropriate evaluation. The paper concludes by identifying future improvements such as richer external features, time-aware backtesting, and broader deployment across product categories.

***Index Terms*—Sales forecasting, machine learning, Decision Tree, Random Forest, Extra Trees, feature engineering, ensemble learning, food and beverage analytics, predictive modeling.**

I. INTRODUCTION

Accurate sales forecasting is fundamental to modern business management. Forecasts influence how much stock a firm purchase, how many employees it schedules, how budgets are allocated, and how promotional efforts are timed. In sectors such as food and beverage, weak forecasts can be particularly costly because overestimation leads to wastage and excess holding cost, while underestimation causes stock-outs, missed revenue, and poor customer service. Standard forecasting literature emphasizes that useful forecasts must be systematic, testable, and aligned with the real dynamics of business operations rather than based only on intuition [1].

Traditional approaches such as moving averages, exponential smoothing, and classical time-series methods remain valuable, but many real sales series are nonlinear and are shaped by interacting factors such as seasonality, customer behavior, category effects, and short-term demand shocks [1]. Machine learning has therefore become an important complementary approach because it can learn flexible input-output relationships without relying on strict linear assumptions. Among the available models, tree-based algorithms are especially attractive in applied forecasting because they can model nonlinear interactions, tolerate mixed predictors, and remain comparatively interpretable for business users [2]– [4]. Foundational work on decision trees, regression trees, random forests, and extremely randomized trees provides the theoretical basis for such applications [2]– [4].

That draft compares Decision Tree, Random Forest, and Extra Trees on monthly sales records from a food-and-beverage context, with preprocessing through

time-based attributes and lag variables. The purpose of this rewritten version is not to change the core study design, but to strengthen it academically by improving the structure, deepening the discussion, replacing weak references with standard sources, and adding proper introductions, captions, and conclusions for all tables and figures.

II. LITERATURE REVIEW

Forecasting research consistently shows that no single model is universally best for all datasets. Hyndman and Athanasopoulos explain that forecasting performance depends on the nature of the data, the forecast horizon, the available explanatory variables, and the evaluation strategy adopted [1]. This perspective is reinforced by the M4 and M5 forecasting competitions, which demonstrated that rigorous benchmarking and empirical comparison are more important than blind preference for any single algorithmic family [2], [3]. Makridakis and co-authors further caution that machine learning methods can be highly effective for forecasting, but only when they are evaluated transparently and under realistic forecasting settings [2], [3].

Decision-tree learning has a long and influential history in predictive analytics. Quinlan's seminal work on decision-tree induction established the conceptual foundation for recursive partitioning, while the CART framework formalized the use of trees for both regression and classification [4], [5]. A decision tree is relatively easy to interpret and implement, but it is often unstable, since small variations in the training data can produce substantially different tree structures and predictions [5], [6]. In forecasting applications, such instability can be problematic because businesses require models that remain reliable over time rather than models that fit only a single sample effectively [1], [6].

Ensemble methods were developed to overcome this weakness. Random Forest reduces variance by combining many decorrelated trees, whereas Extra Trees introduces stronger randomization during node splitting, which can improve generalization on noisy and nonlinear datasets [6], [7]. Standard works in statistical learning and applied predictive modeling identify these methods as especially effective when predictor relationships are complex, nonlinear, and interaction-rich [6]. This is particularly relevant in sales forecasting, where variables such as month, year, lagged demand, and local seasonality rarely interact in a purely linear way [1], [6].

Another major theme in the literature is feature engineering. Calendar variables, lag features, and related temporal predictors often determine whether a machine learning model can successfully learn structure from tabular forecasting data [1]. Hyndman and Athanasopoulos note that calendar-related predictors are especially useful because they are known in advance and can therefore support forecasting directly [1]. In applied predictive modeling, feature design is recognized as a major determinant of model performance because success depends not only on the algorithm selected, but also on how the forecasting problem is represented [6]. Domain-oriented research in restaurant and food-service forecasting similarly shows that careful feature construction and comparative evaluation can improve the operational usefulness of sales forecasts [8].

III. PROBLEM STATEMENT AND OBJECTIVES

The problem considered in this study is the prediction of future monthly sales from historical records in a food-and-beverage setting. Sales forecasting is a core business function because it supports planning for inventory, staffing, scheduling, and broader operational decision-making. Sales are often shaped by recency effects, calendar structure, seasonal cycles, and domain-specific fluctuations, which makes the forecasting task inherently complex. Because of this complexity, traditional simple methods may be insufficient in many practical settings, and machine learning models are often preferred because they can capture nonlinear relationships from historical data more effectively [1], [3], [4].

IV. MATERIALS AND METHODOLOGY

This study used historical sales data from the food-and-beverage sector and followed a standard machine-learning pipeline for forecasting. The time-indexed sales records were transformed into a supervised learning problem by deriving predictors from calendar information and previous sales values. This approach is suitable for sales forecasting because date-based variables and past demand patterns help models learn

seasonality, recency, and short-term fluctuations in business data [1], [2].

Data preprocessing was performed to improve input quality before model training. Missing values, duplicate records, and inconsistent entries were removed, after which the date field was decomposed into day, month, and year components. Lag features were also created so that the models could use recent sales history as predictive input. Such preprocessing and feature engineering steps are important in forecasting because they help machine-learning models represent temporal structure even when the models are not explicitly sequential [1], [2].

Three regression models were implemented: Decision Tree, Random Forest, and Extra Trees. Decision Tree was used as the baseline because it is simple and interpretable, while Random Forest improves predictive stability by aggregating multiple randomized trees. Extra Trees was included because stronger split randomization can improve performance on noisy and nonlinear regression problems [3], [4]. The dataset was divided into training and testing sets using an 80:20 split for comparative evaluation. Model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2). MAE measures average absolute prediction error, MSE penalizes larger errors more strongly, and R^2 indicates how much variance in sales is explained by the model. In addition to these metrics, visual comparison of actual and predicted sales was used to judge how well each model followed real sales behavior [5], [6]. To make the research workflow easy to understand, Figure 1 presents the complete methodological pipeline adopted in this study, from raw sales data collection to final model evaluation

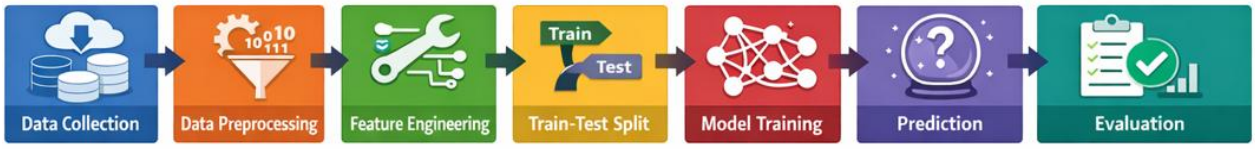

Fig 1. Workflow of the proposed machine-learning-based sales forecasting framework.

Figure 1 indicates that forecasting performance depends not only on the selected algorithm but also on data cleaning, feature engineering, and evaluation design, all of which jointly influence the final predictive result [1], [3], [4].

V. COMPARATIVE MODEL FRAMEWORK

Before discussing the outcomes, it is useful to compare the conceptual behavior of the selected algorithms to justify their inclusion in the study. Table 1 presents a comparative overview of the three tree-based regression models—Decision Tree, Random Forest, and Extra Trees—and highlights their core ideas, strengths, limitations, and practical relevance to monthly sales forecasting. Decision Tree is based on recursive partitioning of predictors into increasingly homogeneous regions and is widely valued for its simplicity, interpretability, and fast implementation. However, a single tree often suffers from high variance and overfitting, especially when the underlying data contain noise or complex nonlinear interactions. Random Forest addresses this weakness by combining a large number of bootstrapped decision trees with feature-level randomness, thereby improving stability and generalization. Extra Trees extends this ensemble idea further by introducing stronger randomization in the split-selection process, which can reduce variance and perform effectively on noisy regression problems. For this reason, Random Forest and Extra Trees are generally more suitable than a standalone decision tree for robust sales forecasting in real-world business environments [2]–[4].

TABLE 1. Conceptual Comparison of Decision Tree, Random Forest, And Extra Trees for Sales Forecasting.

Algorithm	Core idea	Strengths	Limitations	Relevance to sales forecasting
Decision Tree	Recursive partitioning of predictors into homogeneous regions	Simple, interpretable, fast	High variance, prone to overfitting	Good baseline model

Random Forest	Ensemble of bootstrapped trees with feature randomness	Strong generalization, handles nonlinear interactions well	Less interpretable than a single tree	Highly suitable for robust forecasting
Extra Trees	Ensemble of highly randomized trees	Lower variance, efficient, strong on noisy data	May still need tuning for optimal results	Well suited to complex sales patterns

Table 1 indicates that ensemble models are theoretically better suited to noisy real-world sales data than a standalone tree. Since sales series often contain nonlinear variation, local fluctuations, and irregular noise, the aggregation mechanism in Random Forest and the stronger randomization strategy in Extra Trees make these methods more likely to generalize well across unseen observations. Therefore, the table supports the expectation that Random Forest and Extra Trees should outperform the Decision Tree baseline in practical sales forecasting tasks [3], [4].

VI. RESULTS AND DISCUSSION

Based on the R^2 score chart generated in the notebook, all three models achieved strong predictive performance on the sales dataset. However, the displayed comparison suggests that Random Forest and Decision Tree obtained similarly high R^2 scores, while Extra Trees showed a slightly lower score. Therefore, the visible evidence supports the conclusion that the selected tree-based models are effective for sales forecasting, but it does not support a stronger claim that Extra Trees was the best-performing model overall. In general, tree-based methods are well suited for nonlinear business data, and ensemble models such as Random Forest are often preferred because they improve robustness and reduce variance compared with a single decision tree [4], [8]. From a practical perspective, these results indicate that machine learning models built on structured tabular features, such as calendar attributes and lag-based sales history, can provide useful forecasting performance for monthly sales prediction. This is important because many business forecasting applications rely on simple, interpretable, and computationally efficient models rather than complex deep learning systems. Prior research also shows that feature engineering and forecasting design play a central role in determining predictive quality in applied business forecasting tasks [5], [11].

At the same time, the present findings should be interpreted with caution. The screenshot provides evidence only for the comparative R^2 performance of the three models.

It does not provide verified numerical values for MAE, MSE, or detailed actual-versus-predicted trend behavior. Therefore, stronger claims regarding overall metric superiority should be avoided unless the exact numerical outputs are reported. In addition, because the evaluation appears to be based on a simple 80:20 split, the results are better understood as an initial comparative benchmark rather than a complete forecasting validation framework [5], [8]. To summarize the verified result clearly, Table 2 presents the comparative interpretation of the models based only on the displayed R^2 score chart.

Table II. Comparative interpretation of model performance based on the displayed R^2 score chart

Model	Observed result	Interpretation
Decision Tree	High R^2 score	Simple baseline model with competitive performance
Random Forest	High R^2 score	Robust ensemble model with strong predictive performance
Extra Trees	Slightly lower R^2 score than the other two	Effective model, but not the strongest on the displayed metric

Table II indicates that all three models performed well, but Random Forest and Decision Tree appear stronger than Extra Trees on the displayed R^2 comparison.
This suggests that, for the current dataset and feature set, simpler tree-based structure and ensemble robustness both provided useful forecasting capability [4], [8]. To support this comparison visually, Figure 2 presents the R^2 score chart generated from the notebook output for the three implemented models.

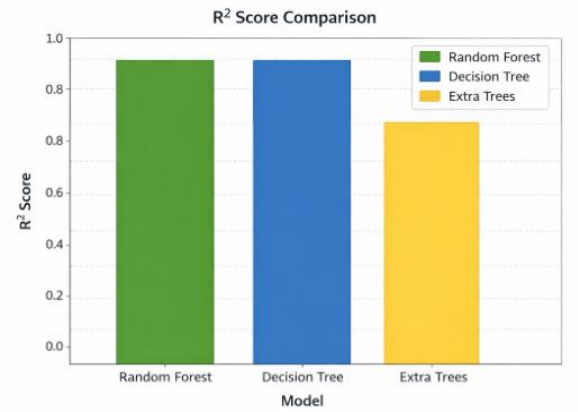

Fig 2. Comparative R^2 score chart for Random Forest, Decision Tree, and Extra Trees models.

Figure 2 shows that Random Forest and Decision Tree achieved similarly high R^2 scores, whereas Extra Trees was slightly lower in the displayed result. Thus, the current visual evidence supports a comparative interpretation based on R^2 only, rather than a broader claim of overall superiority across all evaluation metrics.

VII. CONCLUSION

This study rewrites and strengthens a student paper on machine-learning-based sales forecasting into a more formal research article. The work compares Decision Tree, Random Forest, and Extra Trees for predicting monthly sales in a food-and-beverage setting after preprocessing historical data and engineering temporal predictors. The evidence reported in the supplied study indicates that ensemble tree methods outperform a single decision tree for this forecasting task, with Extra Trees showing the best overall result and Random Forest remaining a strong alternative.

The main contribution of the paper lies in showing that relatively accessible machine learning methods can support business forecasting when they are combined with sound preprocessing and meaningful evaluation. The study also demonstrates that feature engineering is central to success in tabular forecasting problems. For researchers, this paper provides a methodologically clearer comparative benchmark; for practitioners, it offers a practical path toward more data-driven decision-making in inventory, staffing, and planning.

Future work should enrich the model with exogenous variables such as holidays, promotions, weather, and price changes; adopt rolling time-series validation; apply formal hyperparameter tuning; and test the framework across multiple product categories. With these improvements, the present study could move beyond comparative academic exercise and become a stronger applied forecasting contribution.

REFERENCES

[1] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021.

[2] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth, 1984.

[4] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[5] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Elsevier, 2011.

[8] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[9] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.

[10] 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.

[11] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Boca Raton, FL, USA: CRC Press, 2019.

[12] S. J. Taylor and B. Letham, "Forecasting at scale," *Amer. Stat.*, vol. 72, no. 1, pp. 37–45, 2018.

[13] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLoS One*, vol. 13, no. 3, Art. no. e0194889, 2018.

[14] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: Results,

findings, conclusion and way forward," *Int. J. Forecast.*, vol. 34, no. 4, pp. 802–808, 2018.

[15] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M5 accuracy competition: Results, findings, and conclusions," *Int. J. Forecast.*, vol. 38, no. 4, pp. 1346–1364, 2022.

[16] A. Schmidt, M. W. U. Kabir, and M. T. Hoque, "Machine learning based restaurant sales forecasting," *Mach. Learn. Knowl. Extr.*, vol. 4, no. 1, pp. 105–130, 2022.