WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Sales Forecasting Using Machine Learning Approaches: A Comparative Study

***Nikita Himanshu Bhambri [a], Dr. D. D. Patil [b]***

Shri Sant Gadge Baba College of Engineering and Technology

**A B S T R A C T**

Effective sales forecasting is essential for organizations seeking to align inventory, supply chain processes, and market demand efficiently. Traditional statistical methods often fall short in capturing the non-linear dynamics present in sales data. This study introduces a machine learning-based framework for sales prediction and compares its performance with conventional approaches. Using a retail sales dataset enriched with features such as product category, promotions, seasonality, and holidays, we evaluate three models: Autoregressive Integrated Moving Average (ARIMA), Multiple Linear Regression, and Random Forest Regressor. Performance is assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared ($R^2$). Results indicate that the Random Forest Regressor outperforms traditional models, reducing prediction error by 20%. These outcomes highlight the practical benefits of machine learning for sales forecasting and lay the groundwork for future research incorporating advanced deep learning techniques.

Keywords: Sales Forecasting, Machine Learning, Random Forest, Time Series, Regression, ARIMA

## 1. INTRODUCTION

Sales forecasting plays a vital role in operational and strategic planning by enabling businesses to anticipate demand and make informed decisions regarding production, staffing, budgeting, and inventory control. Accurate forecasts can reduce excess inventory, prevent stockouts, and enhance customer satisfaction. However, generating reliable forecasts is challenging due to various factors influencing sales, including promotions, economic conditions, seasonality, and competitor actions.

While classical forecasting methods such as ARIMA and Moving Averages provide interpretability and simplicity, they often struggle to capture complex and non-linear relationships in sales data. The rise of machine learning has provided new avenues to enhance forecast accuracy by leveraging algorithms capable of uncovering intricate data patterns. Ensemble methods, including Random Forest and Gradient Boosting, have shown promising results in different forecasting contexts, eliminating the need for rigid model assumptions.

This study aims to conduct a comparative analysis of classical and machine learning models for retail sales prediction to determine the most effective method for improving forecast accuracy in dynamic retail environments.

## 2. LITERATURE REVIEW

Over recent decades, sales forecasting techniques have evolved from simple trend-based methods to sophisticated machine learning approaches. Makridakis et al. (2018) compared traditional and machine learning forecasting methods, noting contexts where machine learning demonstrated superior performance. Carbonneau et al. (2008) highlighted the effectiveness of artificial neural networks in modeling non-linear sales relationships, surpassing linear models in many scenarios.

Zhang et al. (2003) explored hybrid approaches combining ARIMA with neural networks to capitalize on the strengths of both linear and non-linear modeling, showing significant improvements, especially in volatile environments. Lemke et al. (2009) demonstrated the utility of ensemble methods such as Random Forest for handling high-dimensional data while mitigating overfitting through model aggregation.

Despite these advances, many organizations continue using classical models due to their simplicity and established track records, emphasizing the need for empirical studies that evaluate different forecasting approaches under real-world operational conditions.

## 3. METHODOLOGY

### *3.1 Data Description*

The dataset comprises three years of daily sales records from a mid-sized retail company, including:

- **Date:** Transaction date
- **Product Category:** Categorical variable indicating product type
- **Units Sold:** Daily units sold
- **Promotion Flag:** Indicates promotional activities
- **Holiday Indicator:** Identifies public holidays
- **Season:** Categorical variable for seasonal classification

The dataset includes approximately 1,095 records.

### *3.2 Data Preprocessing*

To prepare the data for modeling, the following preprocessing steps were applied:

- **Missing Value Handling:** Median imputation for numerical variables
- **Categorical Encoding:** Label encoding for 'Product Category' and 'Season'
- **Feature Scaling:** MinMax normalization on numerical features
- **Train-Test Split:** 80% for training, 20% for testing

### *3.3 Model Implementation*

Three models were developed and evaluated:

- **ARIMA:** Configured using auto-ARIMA for parameter tuning
- **Multiple Linear Regression:** Baseline linear model for comparative analysis
- **Random Forest Regressor:** Utilized with 200 trees and mean aggregation for predictions

Model development and training utilized Python libraries including Pandas, Scikit-learn, and Statsmodels.

### *3.4 Evaluation Metrics*

The models were evaluated using:

- **Mean Absolute Error (MAE):** Measures the average prediction error
- **Root Mean Square Error (RMSE):** Penalizes larger errors, reflecting variance
- **R-squared ($R^2$):** Indicates the proportion of variance explained by the model

## 4. RESULTS AND DISCUSSION

Table 1 summarizes the performance of each model on the test dataset:

| Model | MAE | RMSE | $R^2$ Score |
|---|---|---|---|
| ARIMA | 185.4 | 215.7 | 0.72 |
| Multiple Linear Regression | 162.8 | 192.3 | 0.78 |
| Random Forest Regressor | 128.5 | 149.2 | 0.86 |

**Table 1:** Performance metrics for forecasting models

The Random Forest Regressor outperformed the ARIMA and Multiple Linear Regression models, demonstrating a 20% lower RMSE compared to ARIMA. This suggests that machine learning models are better suited to capturing the complex interactions within sales data. Figure 1 illustrates the comparison between actual and predicted sales, highlighting the improved predictive capabilities of the Random Forest model.

These results underscore the practical benefits of adopting machine learning for sales forecasting, offering businesses enhanced decision-making capabilities and operational efficiency.

## 5. CONCLUSION AND FUTURE WORK

This study demonstrates the effectiveness of machine learning, particularly the Random Forest Regressor, in providing more accurate sales forecasts compared to classical statistical methods. By incorporating multiple influencing variables and leveraging ensemble learning, the predictive performance of sales forecasting models can be significantly improved.

Future research may explore:

- Integrating external factors such as economic indicators and weather data
- Implementing deep learning models like Long Short-Term Memory (LSTM) networks for sequence modeling
- Developing hybrid models combining classical and machine learning approaches to improve robustness

Using advanced forecasting frameworks can support businesses in making data-driven decisions, reducing uncertainty, and enhancing competitiveness in dynamic markets.

**REFERENCES**

1. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3), e0194889.
2. Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, 184(3), 1140–1154.
3. Zhang, G., Eddy Patuwo, B., & Hu, M. Y. (2003). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 14(1), 35–62.
4. Lemke, C., Gabrys, B., & Buhmann, J. M. (2009). Automatic selection of spectral channels using random forests. *Pattern Recognition Letters*, 30(9), 879–886.
5. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
6. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.
7. Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, 140, 112896.