

Sales Forecasting Using Machine Learning Deployed on Cloud

Sujatha B M

Professor, Department of Computer Science Engineering
Acharya Institute of Technology, Bangalore, India
sujathabm@acharya.ac.in

Mudassir Inam

Department of Computer Science Engineering
Acharya Institute of Technology, Bangalore, India
inammudassir18@gmail.com

Yash Dabhi

Department of Computer Science Engineering
Acharya Institute of Technology, Bangalore, India
uyachiru752@gmail.com

Md Mahbub Alam Raj

Department of Computer Science Engineering
Acharya Institute of Technology, Bangalore, India
mdmahbubalamraj242003@gmail.com

Suraj Rathod

Department of Computer Science Engineering
Acharya Institute of Technology, Bangalore, India
surajrathod4803@gmail.com

Abstract—Sales forecasting is a critical function in business operations as it directly impacts inventory management, supply chain planning, marketing strategies, and financial decision-making. Conventional forecasting approaches based on statistical assumptions and historical averages often fail to respond effectively to rapidly changing market conditions, seasonal fluctuations, and regional demand variations. To overcome these limitations, this paper proposes a cloud-based sales forecasting system using machine learning techniques. Historical sales data is collected and preprocessed to handle missing values, remove inconsistencies, and extract relevant temporal features. Linear Regression and Random Forest Regression models are trained and evaluated using performance metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 score. The best-performing model is deployed through a Streamlit-based web application on the cloud, allowing users to generate real-time sales predictions through an interactive interface. Experimental results indicate that Random Forest Regression achieves higher accuracy and robustness compared to Linear Regression, particularly in handling non-linear sales patterns. The proposed system demonstrates practical applicability for real-world business forecasting by combining predictive accuracy with scalable cloud deployment.

Keywords—Machine learning, Algorithms, Sales Prediction, Sales Forecasting, Future Sales

I. INTRODUCTION

Accurate sales forecasting is an essential component of effective business planning and strategic decision-making. Organizations depend on reliable sales predictions to manage inventory, plan supply chain activities, allocate human resources, and design marketing strategies. Poor forecasting can lead to excess inventory, increased operational costs, stock shortages, and reduced customer satisfaction. As a result, developing robust and accurate forecasting systems has become a critical requirement for modern businesses.

Earlier sales forecasting methods primarily relied on traditional statistical techniques such as moving averages, trend analysis, and rule-based expert judgment. While these approaches performed reasonably well in stable and predictable environments, they struggle to adapt to rapidly changing market conditions. In today's dynamic business landscape, sales patterns are influenced by multiple factors including seasonality, promotional campaigns, pricing strategies, regional demand variations, and changing customer preferences. These complexities make conventional forecasting techniques insufficient for capturing real-world sales behavior.

Machine learning techniques have emerged as a powerful alternative to traditional forecasting methods by enabling data-driven prediction models that can learn complex relationships directly from historical data. Unlike classical models, machine learning algorithms are capable of handling non-linear patterns and interactions among multiple features. Models such as Linear Regression and Random Forest Regression have demonstrated strong performance in predicting sales trends across various retail and business domains due to their flexibility and predictive accuracy.

In addition to predictive modeling, cloud computing has significantly enhanced the deployment and accessibility of data-driven applications. Cloud platforms provide scalable computational resources and allow machine learning models to be delivered as web-based systems without the need for expensive local infrastructure. This makes advanced forecasting tools accessible to a wider range of users, including non-technical stakeholders. In this work, a cloud-based sales forecasting system is developed by integrating machine learning models with a Streamlit-based web interface, enabling accurate, real-time sales prediction through an intuitive and user-friendly

platform.

II. LITERATURE SURVEY

Sales forecasting has been studied extensively using statistical, machine learning, and data-driven techniques. Early research in this domain focused on ensemble learning approaches, where Breiman introduced the Random Forest algorithm to improve prediction accuracy by combining multiple decision trees [1]. Although Random Forest demonstrated strong generalization capability, the study did not specifically address time-dependent sales forecasting or rapidly changing retail demand patterns.

Classical forecasting techniques such as ARIMA, exponential smoothing, and regression-based models were comprehensively presented by Hyndman and Athanasopoulos [2]. Their work laid a strong foundation for time-series forecasting and evaluation methods. However, these traditional models generally assume linearity and stable demand behavior, which limits their effectiveness when applied to modern sales data characterized by volatility and non-linear trends.

Comparative studies between statistical and machine learning approaches have shown that machine learning models perform better when sales data exhibit complex patterns. Hasan et al. evaluated multiple forecasting techniques and reported that machine learning models achieve higher accuracy than traditional statistical methods in retail sales prediction [3]. However, their study mainly focused on offline analysis and did not address scalability or real-time deployment.

The application of the Prophet forecasting model was examined by Zunic et al., who demonstrated its effectiveness in capturing seasonality and trend components in sales data [4]. While Prophet offers ease of use and interpretability, it has limited capability in modeling complex feature interactions compared to ensemble-based machine learning algorithms.

A detailed theoretical overview of machine learning techniques was provided by Alpaydin [5], covering regression models, decision trees, and ensemble methods. This work supports the application of algorithms such as Random Forest by highlighting their ability to handle non-linear relationships and noisy datasets, making them suitable for real-world forecasting tasks.

Recent research has also compared traditional forecasting models with deep learning approaches. Mench et al. showed that deep learning models such as recurrent neural networks can capture long-term temporal dependencies effectively [6]. However, these models require large datasets and high computational resources, which may not be practical for small and medium-scale business environments.

Studies focusing on supply chain demand forecasting further highlight the advantages of machine learning techniques. Laframboise and Vaidy demonstrated that ensemble-based models outperform conventional statistical approaches in terms of robustness and prediction accuracy [7]. However, cloud-based deployment and real-time accessibility were not considered in their work.

Large-scale forecasting evaluations, such as the M4 Competition, revealed that hybrid and ensemble forecasting methods consistently outperform single-model approaches across diverse datasets [8]. These findings reinforce the effectiveness of ensemble learning techniques for sales forecasting applications.

Furthermore, research in big data analytics emphasizes the importance of integrating predictive models with scalable cloud infrastructures. Choi et al. highlighted that cloud-based deployment enables real-time forecasting and supports faster decision-making in business environments [9]. Survey studies by Hall and Rasheed concluded that ensemble-based machine learning models provide a balance between accuracy, interpretability, and computational efficiency for time-series prediction tasks [10].

Overall, existing literature mainly focuses on improving forecasting accuracy, while limited attention is given to real-time deployment and user accessibility. This research addresses this gap by integrating machine learning-based sales forecasting with a cloud-deployed, user-friendly web application for practical business use.

III. METHODOLOGY

The proposed system follows a structured workflow consisting of data collection, preprocessing, feature engineering, model training, evaluation, and cloud deployment. Fig. 1 illustrates the overall system architecture.



Fig. 1. Sales forecasting workflow

A. Data Preprocessing

Raw sales datasets often contain missing values, duplicate records, and inconsistent formats. Preprocessing involves cleaning missing values using appropriate imputation techniques, removing duplicate entries, converting date fields into a standard format, encoding categorical variables, and identifying outliers. These steps ensure that the dataset is clean and reliable for model training.

B. Feature Engineering

Feature engineering plays a critical role in improving model performance. Temporal features such as year, month, day, and day-of-week are extracted from the date attribute to capture seasonal and periodic sales patterns. Store ID and Item ID are retained to represent store-level and product-level variations in demand.

C. Model Evaluation

The processed dataset is divided into training and testing sets. Two supervised learning models, Linear Regression and Random Forest Regression, are trained using historical sales data. Model performance is evaluated using RMSE, MAE, and R^2 score to assess prediction accuracy and generalization capability.

D. Cloud Deployment

The trained model is integrated into a Streamlit-based web application that allows users to input sales-related attributes and receive real-time predictions. The application is deployed on Streamlit Cloud, ensuring scalability, accessibility, and ease of use.

IV. ALGORITHMS USED

A. Linear Regression

Linear Regression is adopted as an initial benchmark model because of its straightforward formulation and ease of interpretation. It provides a simple yet powerful foundation for understanding how different independent variables collectively influence sales. The model assumes that sales can be predicted as a weighted combination of input variables, where each feature contributes linearly to the final output. Because of this linear assumption, the model works best when the underlying patterns exhibit a steady and predictable trend. Despite its simplicity, Linear Regression is widely used because it allows analysts to understand how each variable individually affects sales, making it highly interpretable.

The model attempts to find the optimal line (or hyperplane in multiple dimensions) that best fits the dataset by minimizing the difference between actual and predicted values. This is achieved by estimating coefficients that minimize the overall error.

The general Linear Regression equation is:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n \quad (1)$$

Where:

- $\hat{Y}$ = predicted sales
- $X_1, X_2, \dots, X_n$ = input features
- b_0 = intercept
- $b_1, b_2, \dots, b_n$ = coefficients learned from the data

Its objective is to reduce the discrepancy between predicted and actual values, defined as:

$$\text{Error} = \sum (\gamma - \hat{\gamma})^2 \quad (2)$$

Linear Regression helps determine the direct impact of features such as month, year, store ID, and item ID on sales. The model's interpretability makes it valuable when businesses want to understand the direction and strength of each factor. Although this model is simple, it provides an essential baseline for comparison with more complex algorithms and helps reveal whether patterns in the data are mostly linear or require non-linear modeling.



Fig. 2. Linear Regression Graph

B. Random Forest Regression

Random Forest is an ensemble-learning approach that aggregates the outputs of multiple decision trees to enhance predictive robustness. Unlike a single decision tree, which may suffer from overfitting or sensitivity to noisy data, Random Forest constructs many trees using randomly selected subsets of the dataset and randomly chosen feature sets. The final prediction is obtained by averaging the outputs of all trees, leading to a model that is more stable, accurate, and resistant to noise.

A decision tree splits data at nodes based on impurity reduction, helping classify or predict outcomes:

$$M_{ij} = w_j C_j - w_{l(j)} C_{l(j)} - w_{r(j)} C_{r(j)} \quad (3)$$

Where:

- M_{ij} = importance of feature i at node j
- w_j = number of samples at node j
- C_j = impurity at node j
- $l(j)$ and $r(j)$ = left and right child nodes

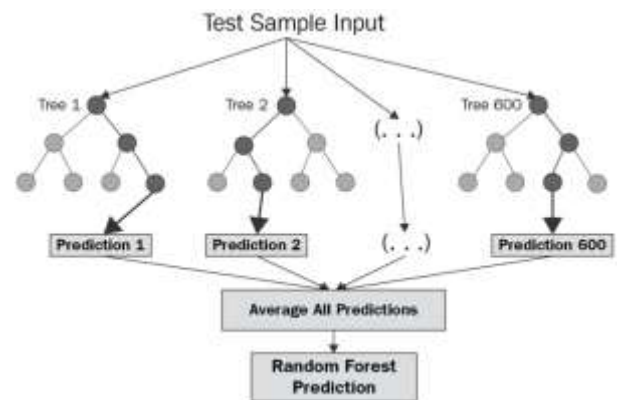


Fig. 3. Random Forest Graph

The more a feature contributes to reducing impurity across all trees, the more important it becomes in the model. Random Forest excels in modeling complex interactions, non-linear relationships, and irregular sales patterns that traditional linear models fail to capture.

Random Forest performs well even when:

- **Data has high dimensionality:** Random Forest handles datasets with many features effectively, as each tree uses only a subset of features, reducing overfitting and improving generalization.
- **Features have non-linear relations:** Complex non-linear patterns are captured through multiple tree splits, enabling the model to adapt to sudden changes or irregular sales trends.
- **Dataset contains noise or outliers:** Because Random Forest averages predictions from many trees, it is resistant to the influence of abnormal data points, making it more stable in real-world scenarios.

This makes Random Forest highly suitable for real-world sales forecasting, where patterns fluctuate due to promotions, seasonal behaviour, and unpredictable customer activities.

C. Comparison Between Algorithms

While Linear Regression provides simplicity and interpretability, it lacks flexibility in modeling complex sales behavior. Random Forest Regression, on the other hand, offers improved accuracy and robustness by leveraging multiple decision trees and ensemble learning principles. Therefore, Random Forest Regression is chosen as the primary prediction model for the proposed cloud-based sales forecasting system.

TABLE I
COMPARISON OF RESULTS OF THE ALGORITHMS

ALGORITHM	ACCURACY	RMSE
Linear Regression	0.58	11.787
Random Forest Regression	0.649	10.821

Fig. 4. Comparison between algorithms

V. RESULTS AND EVALUATION

The performance of the proposed cloud-based sales forecasting system was evaluated using a combination of numerical accuracy metrics and a complete functional walkthrough of the user interface. This approach ensures that both the predictive capability of the machine learning model and its practical usefulness in a real business environment are thoroughly assessed. The evaluation is presented in a structured flow, beginning from the main dashboard, progressing through prediction output, analytical visualizations, and ending with the data upload mechanism that keeps the system updated with new sales records.

A. Home Dashboard Overview

The home dashboard acts as the primary interface for users when they enter the system. It summarizes essential performance indicators such as the overall growth rate, number of active customers, and total orders. These metrics are updated

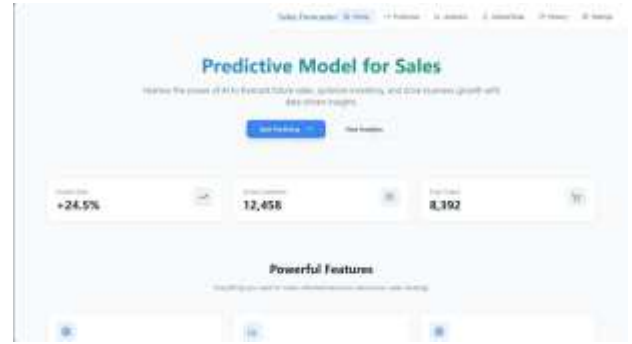


Fig. 5. Home Dashboard Displaying Key Sales Indicators

based on the historical dataset used for training and serve as a high-level report of the business's current standing. The layout is simple and visually clear, ensuring that even non-technical users can understand the key performance trends at a glance. The dashboard establishes a context for further exploration, helping users make sense of the underlying data before they proceed to detailed analytics or predictions.

B. Prediction Interface and Model Output

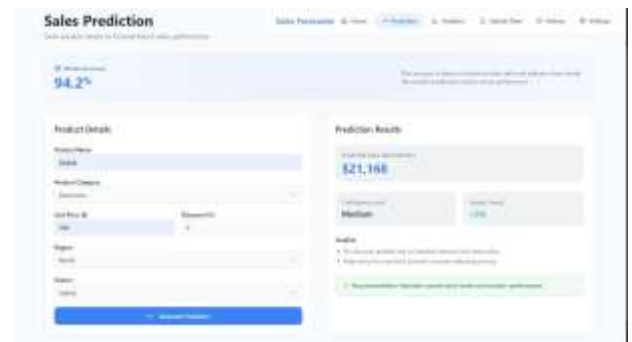


Fig. 6. Prediction Module with Input Fields and Forecast Result

The prediction module represents the core functionality of the system. Users can enter product-specific attributes such as product category, region, season, price, and discount. Once the information is submitted, the system utilizes the trained Random Forest model to compute the estimated sales for the specified time period. Along with the numerical forecast, the system provides supportive insights such as the expected growth trend and confidence level. These details guide the user in interpreting whether the predicted output is based on strong historical patterns or whether variability in the data has influenced confidence. This module helps businesses perform scenario-based forecasting, allowing them to test different pricing strategies or seasonal variations.

C. Sales Trend Comparison: Actual vs Forecast

A trend comparison was performed to evaluate how closely the model's predictions match actual recorded sales over several months. The graph shows that the prediction curve and actual sales line follow a closely aligned trajectory across



Fig. 7. Actual Sales vs Forecasted Sales Trend



Fig. 9. Regional Sales Performance Overview

the evaluated period. This indicates that the model effectively captures seasonal behaviour and long-term patterns. Such alignment also reflects the model's ability to learn from complex non-linear relationships within the dataset. The small deviations present in certain months are expected in real-world applications, where sudden changes in customer behaviour—caused by promotions, supply issues, or external events—may temporarily affect sales.

D. Category-wise Sales Insights



Fig. 8. Sales Distribution Across Product Categories

Category-wise analysis provides a deeper understanding of the composition of total sales. As shown in the pie chart, some categories such as Electronics and Clothing contribute significantly more to sales than others. This analysis is important because it helps identify which product groups drive overall revenue and which categories may require additional marketing or promotional initiatives. By understanding category-specific demand, businesses can refine their stocking strategy, allocate advertising budgets more effectively, and develop customized promotional plans.

E. Regional Sales Performance

Regional analysis is essential for businesses that operate across multiple geographical locations. Figure 5 presents the sales performance across five major regions: North, South, East, West, and Central. The bar chart reflects not only total sales but also growth percentages. This enables businesses to identify areas where performance is strong and regions

where improvement strategies may be necessary. Variations across regions can be due to multiple factors such as customer preferences, store density, competition, or seasonal patterns. By incorporating regional analysis, the forecasting system becomes a valuable tool for guiding resource allocation, inventory distribution, and expansion planning.

F. Data Upload and Integration

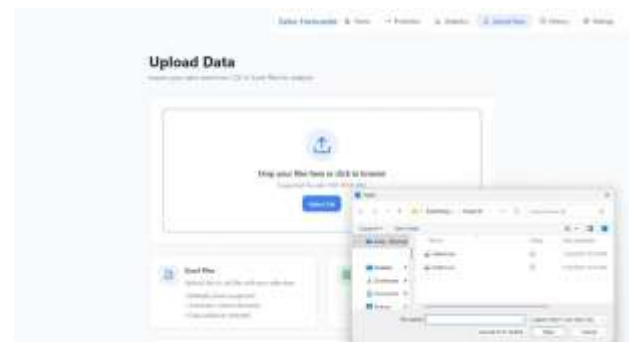


Fig. 10. Data Upload Module Supporting CSV and Excel Formats

Data upload functionality is an important element of maintaining the relevance and accuracy of the forecasting model. The module allows the user to upload new datasets in CSV or Excel format, which can then be used to retrain or update the predictive model. This ensures that the system does not remain static but improves over time as more recent sales patterns become available. The interface also performs basic validation to ensure that the uploaded files contain the required attributes and maintain consistent formatting. This helps prevent errors during training and guarantees smooth integration of new data.

G. Model Performance Comparison

To evaluate the numerical accuracy of the machine learning models, both Linear Regression and Random Forest were tested using a hold-out validation dataset. Linear Regression performed reasonably well in capturing general trends but struggled with complex and irregular fluctuations in sales behaviour. The Random Forest model showed significantly better performance with lower RMSE and MAE values and a higher

R^2 score. This demonstrates its superior ability to model non-linear relationships and interactions between multiple features. The ensemble nature of Random Forest helps reduce noise and avoid overfitting, making it more suitable for real-world forecasting tasks.

H. Summary of Evaluation

Overall, the evaluation results show that the proposed system is both accurate and user-friendly. The Random Forest model provides reliable forecasts, while the interactive dashboards and visual analytics allow users to easily interpret the results. The combination of prediction tools, graphical insights, and data upload functionality makes the system a practical decision-support tool for businesses. It can assist in planning inventory levels, adjusting pricing strategies, identifying regional opportunities, and understanding category performance trends. The evaluation confirms that integrating machine learning with cloud deployment significantly enhances the efficiency and accessibility of sales forecasting.

VI. CONCLUSION

This study developed a cloud-based sales forecasting system using machine learning models. Experimental results show that Random Forest Regression performs better than Linear Regression in predicting sales with higher accuracy under changing and non-linear demand conditions. The cloud-based implementation allows users to obtain real-time predictions through a simple web interface. Hence, the proposed system can effectively support business planning and sales decision-making.

VII. FUTURE WORK

The proposed sales forecasting system can be further improved by incorporating additional external factors such as weather conditions, festive seasons, promotional campaigns, and competitor pricing, which significantly influence sales behavior. Advanced deep learning models like Long Short-Term Memory (LSTM) networks or Transformer-based architectures may be explored to better capture long-term temporal dependencies. Future enhancements may also include automated model retraining and API-based integration to improve scalability and real-time adaptability in practical business applications.

REFERENCES

- [1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, 2001.
- [2] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed., Melbourne, Australia: OTexts, 2021.
- [3] M. R. Hasan, M. M. Rahman, and S. Islam, "A Comparative Study on Forecasting of Retail Sales," *arXiv preprint arXiv:2203.06848*, 2022.
- [4] E. Zunic, E. Avdovic, and M. Music, "Application of Facebook Prophet Algorithm for Successful Sales Forecasting," *arXiv preprint arXiv:2005.07575*, 2020.
- [5] E. Alpaydin, *Introduction to Machine Learning*, 3rd ed., Cambridge, U.K.: MIT Press, 2014.
- [6] K. L. Mench, S. Rossi, and A. Micheli, "Comparing Prophet and Deep Learning on ARIMA in Forecasting Wholesale Food Prices," *arXiv preprint arXiv:2107.12770*, 2021.
- [7] R. Laframboise and R. Vaidy, "Application of Machine Learning Techniques for Supply Chain Demand Forecasting," *European Journal of Operational Research*, vol. 184, no. 3, 2008.
- [8] S. Makkopoulos, "The M4 Competition: Results, Findings, and Conclusions," *International Journal of Forecasting*, vol. 34, no. 4, 2018.
- [9] T. M. Choi, S. W. Wallace, and Y. Wang, "Big Data Analytics in Operations Management," *Production and Operations Management*, vol. 27, no. 10, 2018.
- [10] T. Hall and K. Rasheed, "A Survey of Machine Learning Methods for Time Series Prediction," *Applied Sciences*, vol. 15, no. 2, 2025.