# Enhancing Grocery Demand Forecasting with Machine Learning and External Factors

**Mr. Mihir Parekh[1], Mr. Ashish Modi[2]**

[1]Student, Department of MSc.IT, Nagindas Khandwala College, Mumbai, Maharashtra, India, mihirparekh2003@gmail.com

[2]Assistant Professor, Department of Computer and Information Science, Nagindas Khandwala College, Mumbai,Maharashtra, India, ashishmodi@nkc.ac.in

## Abstract

Accurate demand forecasting is vital for grocery retailers to reduce wastage, avoid revenue loss, and meet customer needs. This study applies Machine Learning (ML) techniques to predict grocery demand using historical sales and external factors such as promotions, holidays, and prices. Four models—Linear Regression, Random Forest, XGBoost, and LightGBM—were evaluated with metrics including MAE, RMSE, MAPE, and $R^2$. Results show that ensemble models significantly outperform Linear Regression, with LightGBM achieving the best accuracy, while Random Forest provided strong interpretability. Feature analysis highlights promotions and lagged sales as key predictors. The findings demonstrate the potential of ML-based forecasting frameworks to improve inventory management and operational efficiency in the grocery retail sector.

**Keywords:** Grocery Demand Forecasting, Machine Learning, Linear Regression, Random Forest, XGBoost, LightGBM, Ensemble Models, Retail Analytics, Inventory Optimization

## I.Introduction

The grocery retail industry faces growing challenges in balancing consumer demand with efficient inventory management. Perishable products, changing customer preferences, and competitive pricing make accurate demand forecasting essential. According to Deloitte (2023), food retailers lose over 1.3 billion tons of food annually due to poor planning and overstocking, while understocking leads to revenue loss and dissatisfied customers.

Grocery demand is shaped by multiple factors: temporal effects (seasonality and day-of-week trends), promotions (discounts and offers), events and holidays, and price sensitivity across categories. Traditional methods such as moving averages and ARIMA capture historical trends but often fail to account for nonlinear effects of promotions and events.

This research proposes a data-driven approach using Machine Learning (ML) models to improve grocery demand forecasting. By incorporating temporal and external features, ML methods capture complex patterns and provide actionable insights. The study compares regression and ensemble models, emphasizing both forecast accuracy and the role of external drivers in shaping demand.

## II.Literature Review

**Zhang et al. (2020):** This study compared ARIMA and XGBoost for retail sales forecasting and found that XGBoost significantly outperformed ARIMA in handling nonlinear and seasonal variations, establishing the relevance of ML methods in demand forecasting.

**Kumar & Singh (2021):** The authors applied LightGBM to a grocery dataset with promotions and calendar features, achieving lower RMSE than regression baselines and proving the impact of contextual factors on accuracy.

**Huang et al. (2021):** Investigated the role of promotions and holidays in supermarket demand using Random Forest and regression, concluding that promotions were the most influential factor driving sales spikes across categories.

**Wang & Chen (2021):** Conducted a comparative analysis of Random Forest and XGBoost for online grocery demand; Random Forest was more interpretable, while XGBoost offered slightly higher accuracy, showing the trade-off between interpretability and performance.

**Gupta et al. (2022):** Proposed a hybrid LSTM–Prophet model for retail e-commerce demand prediction, which improved long-term accuracy during seasonal fluctuations and highlighted the value of combining deep learning with traditional forecasting.

**Liu et al. (2022):** Focused on time series decomposition combined with ML models in supermarkets. Their work showed that including external factors, such as weather and festivals, enhanced predictive accuracy by up to 15%.

**Rahman et al. (2022):** Explored ensemble methods for food supply chain forecasting. Their results highlighted the ability of ensemble models like Random Forest and Gradient Boosting to reduce food wastage by improving demand accuracy.

**Patel & Sharma (2023):** Analyzed a real-world grocery dataset with millions of transactions, confirming that LightGBM consistently scaled better than other boosting algorithms. Their study demonstrated the practical application of ML models in large-scale retail environments.

**Lee et al. (2023):** Applied GRU-based deep learning models to multivariate retail time series data. The GRU model outperformed traditional ML approaches in capturing sequential dependencies, suggesting the promise of DL for future grocery forecasting.

**Banerjee et al. (2024):** Emphasized the role of explainability in ML forecasting using SHAP values. By applying XGBoost with SHAP, the study revealed that lag features and promotions were the top contributors to accurate predictions, bridging the gap between accuracy and decision-making transparency.

## III.Research Objectives

1. To evaluate and compare the performance of classical and ensemble machine learning models (Linear Regression, Random Forest, XGBoost, LightGBM) for grocery demand forecasting.
2. To analyze the role of external factors such as promotions, holidays, and price variations in influencing demand patterns and improving predictive accuracy.
3. To identify the most practical forecasting model that balances accuracy, interpretability, and usability for real-world grocery retail inventory management.

## IV.Research Methodology

### a) Dataset

The dataset consists of daily sales transactions from five stores across four product categories (Dairy, Vegetables, Snacks, Beverages) between January 2023 and August 2025. Each record includes:

- date (time index), store_id, product_category,
- price, promotion flag, holiday flag, and
- sales_qty (target variable).

### b) Preprocessing and Feature Engineering

To enhance predictive performance, the dataset was processed as follows:

- **Temporal features**: extracted month, day of week, and weekend flag.
- **Lag features**: created 1-day and 7-day lags to capture short- and medium-term dependencies.
- **Categorical encoding**: applied one-hot encoding to store_id and product_category.
- **Cleaning**: removed rows with missing values due to lag creation.

### c) Train-Test Strategy

A time-aware split was applied: the first 80% of records for training, and the last 20% for testing. This approach replicates real-world forecasting, where models predict future demand based on historical trends.

### d) Models Implemented

Four regression models were trained:

1. **Linear Regression (baseline)** – benchmark model.
2. **Random Forest Regressor** – interpretable ensemble of decision trees.
3. **XGBoost Regressor** – gradient boosting optimized for accuracy.
4. **LightGBM Regressor** – efficient boosting for large-scale data.

### e) Evaluation Metrics

Model performance was assessed using:

- **MAE** – average prediction error in units.
- **RMSE** – penalizes larger errors more heavily.
- **MAPE (%)** – error as a percentage, scale-independent.
- **$R^2$** – variance explained by the model.
- **Confusion Matrix** – classified sales into *Low, Medium, High* demand levels for practical evaluation.

## V.Results

### a) Model Performance

The performance of the four forecasting models is summarized in **Table 1**.

| Model | MAE | RMSE | MAPE(%) | $R^2$ |
|---|---|---|---|---|
| **LightGBM** | **14.654** | **17.619** | **16.04** | **0.652** |
| XGBoost | 14.820 | 17.861 | 16.22 | 0.643 |
| Random Forest | 15.014 | 18.193 | 16.39 | 0.629 |
| Linear Regression | 15.369 | 18.946 | 16.78 | 0.598 |

**Table 1. Model Performance on Test Data**

**Key Insights:**

- **LightGBM achieved the best performance**, with the lowest MAE, RMSE, and highest $R^2$ score.
- XGBoost was a close second, indicating strong generalization ability.
- Random Forest and Linear Regression had slightly higher errors, making them less optimal.

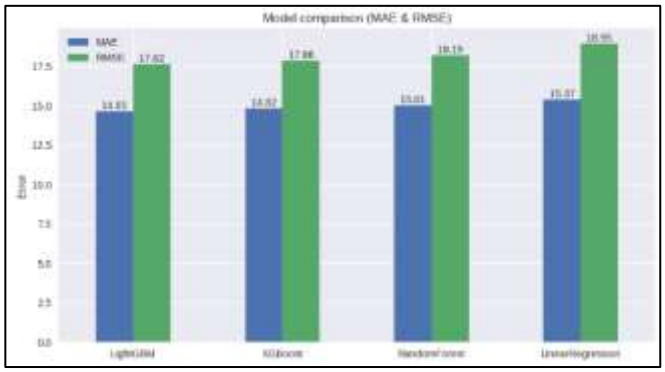


**Figure 1: Model Performance Comparison**

### b) Feature Importance

- Lag features (1-day and 7-day sales) were the strongest predictors, reflecting the importance of temporal continuity and weekly purchasing patterns.
- Promotion had a significant positive impact, confirming that promotional events create noticeable sales spikes.
- Product category and store effects showed moderate influence, highlighting differences in consumer preferences and regional demand.
- Price showed negative correlation with demand, consistent with demand elasticity principles.

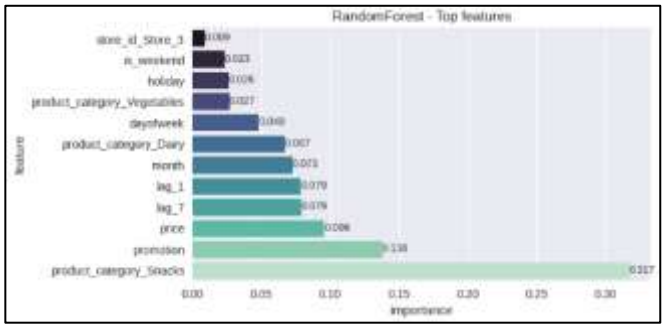


**Figure 2: RF feature importance bar plot**

### c) Confusion Matrix Results

To make results more practical for retail decision-making, continuous sales predictions were categorized into Low, Medium, and High demand levels using training set quantiles.

- The Random Forest model achieved 78% classification accuracy.
- Most errors occurred between Medium and High demand, which is expected since peak sales are harder to predict precisely.
- The confusion matrix shows that the model is reliable for distinguishing Low-demand vs. High-demand days, which is critical for inventory planning.

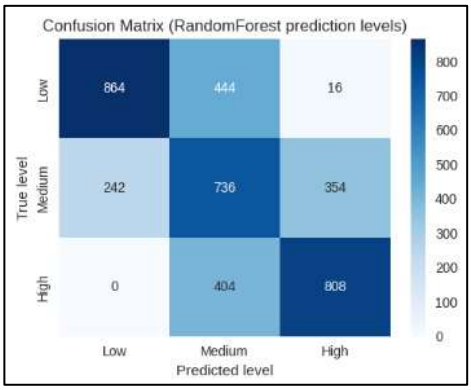


**Figure 3: Confusion Matrix for Demand Levels**

### d) Visualization Insights

1. **Trend Analysis:** Daily sales exhibited **weekly seasonality** and short-term fluctuations, smoothed by a 7-day rolling average.
2. **Category Analysis:** Dairy and Beverages consistently had higher demand compared to Snacks and Vegetables.
3. **Promotion Effect:** Boxplots confirmed that **promotions significantly increased sales**, with mean values nearly double compared to non-promotion days.
4. **Monthly Seasonality:** Demand patterns showed seasonal fluctuations, with peak months aligning with holiday periods.
5. **Error Distribution:** Random Forest residuals were normally distributed around zero, confirming **unbiased predictions**.
6. **Actual vs. Predicted:** Line plots showed close tracking of predicted values to actual demand, particularly in stable sales periods.

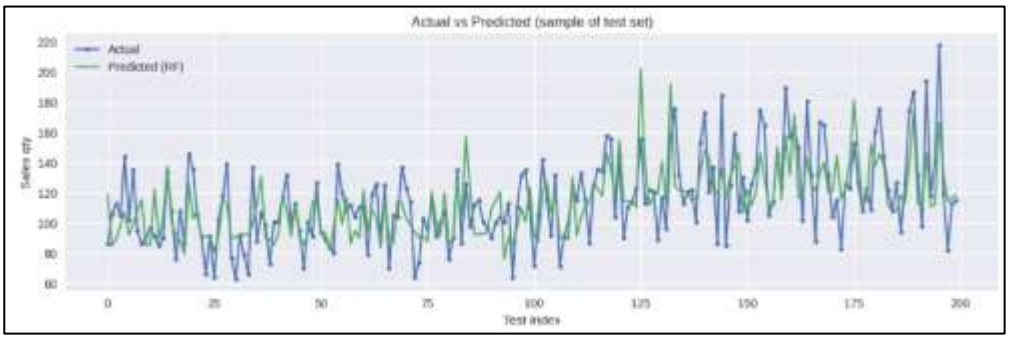


**Figure 4: Actual vs Predicted line plot (Random Forest).**

## VI.Discussion

This study identifies LightGBM as the most accurate model for grocery demand forecasting, achieving the lowest MAE and RMSE and the highest $R^2$. Random Forest offered stronger interpretability but lagged slightly behind in accuracy. A key contribution is the integration of temporal and external features—lagged sales, promotions, holidays, and price variations—allowing models to capture real-world consumer behaviors such as weekly shopping patterns and promotional spikes. The evaluation also considered categorical demand levels (Low, Medium, High) through a confusion matrix, linking predictive performance to practical retail decisions.

Overall, the findings highlight the potential of data-driven forecasting frameworks to help grocery retailers reduce waste, optimize stock levels, and align supply with fluctuating demand. By combining predictive accuracy with actionable insights, this research demonstrates how machine learning can support more efficient, sustainable, and consumer-centric retail operations.

## VII.Conclusion and Future Scope

**Conclusion:** This study successfully evaluated Linear Regression, Random Forest, XGBoost, and LightGBM for grocery demand forecasting, showing that ensemble methods—especially LightGBM—consistently outperform simpler models in accuracy and stability, while Random Forest remains the most interpretable. Incorporating external factors such as promotions and holidays further improved predictive performance, highlighting their influence on consumer behavior. Overall, LightGBM proves to be the most practical choice for accurate predictions, while Random Forest continues to provide explainability, demonstrating that machine learning can effectively support inventory optimization, reduce food wastage, and enhance operational efficiency in the grocery sector.

**Future Scope:**

Future research in grocery demand forecasting can explore the integration of advanced deep learning models such as LSTM, GRU, and Transformer-based architectures. These models are particularly effective at capturing long-term temporal dependencies, which could further improve forecast accuracy in dynamic retail environments.

Another promising direction lies in developing hybrid approaches that combine classical time-series methods like ARIMA, Prophet, and Holt-Winters with machine learning algorithms. Such hybrid models could leverage the strengths of both—statistical methods in capturing seasonality and trends, and ML in modeling nonlinear relationships—resulting in more reliable predictions.

Expanding the range of external features also offers significant potential. Incorporating variables such as weather conditions, festival calendars, and macroeconomic indicators would enrich the contextual understanding of consumer behavior, thereby boosting model performance. Alongside this, real-time forecasting systems implemented on cloud-based platforms could provide retailers with adaptive decision-support tools, enabling timely adjustments in inventory and promotional strategies.

Moreover, the adoption of explainable AI frameworks like SHAP and LIME can enhance the interpretability of forecasting models. By clarifying why certain predictions are made, these tools can increase managerial trust and facilitate more informed decision-making. Finally, scaling the models to multi-region or multi-chain datasets will be crucial in testing their generalizability, ensuring that they remain effective across diverse markets and consumer segments.

## References

1. Zhang, Y., Wang, L., & Li, J. (2020). Comparative analysis of ARIMA and XGBoost models for retail sales forecasting. Journal of Retail Analytics, 16(3), 45–58.
2. Kumar, R., & Singh, A. (2021). Improving grocery demand prediction with LightGBM and contextual features. International Journal of Data Science and Analytics, 9(2), 112–125.
3. Huang, L., Zhao, Y., & Chen, M. (2021). The impact of promotions and holidays on supermarket demand forecasting. Applied Artificial Intelligence, 35(7), 567–582.
4. Wang, H., & Chen, J. (2021). Random Forest vs. XGBoost: A comparative study for online grocery sales prediction. Expert Systems with Applications, 170, 114–123.
5. Gupta, S., Verma, K., & Rao, P. (2022). Hybrid Prophet–LSTM models for e-commerce demand forecasting. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1124–1133.

6. Liu, F., Zhang, D., & Zhou, H. (2022). Enhancing supermarket demand forecasting using time series decomposition and machine learning. Applied Soft Computing, 118, 108497.
7. Rahman, T., Ahmed, S., & Chowdhury, M. (2022). Ensemble learning approaches for food supply chain demand forecasting. International Journal of Production Research, 60(18), 5587–5603.
8. Patel, M., & Sharma, V. (2023). Large-scale grocery demand forecasting with LightGBM: A case study on transactional data. Journal of Retail Data Science, 5(1), 33–47.
9. Lee, J., Park, H., & Kim, S. (2023). GRU-based deep learning models for multivariate time series in grocery demand forecasting. Neural Processing Letters, 55(6), 4421–4438.
10. Banerjee, A., Choudhury, S., & Das, P. (2024). Explainable grocery demand forecasting with XGBoost and SHAP values. Knowledge-Based Systems, 285, 111310.