Optimizing Retail Demand Forecasting with Advanced Machine Learning Techniques

Chanicha Moryadee[1], Komson Sommanawat [2], Mano Prachayapipat [3], and Thun Chaithon[4]

Email: chanicha.mo@ssru.ac.th[1], komson.so@ssru.ac.th[2],
mano.pr @ssru.ac.th[3], and thun.ch@ssru.ac.th[4]

[1,2,3,4]College of Logistics and Supply Chain, Suan Sunandha Rajabhat University, Bangkok, Thailand

*Corresponding author

Abstract

Accurate demand forecasting is critical for retail businesses to optimize inventory management, reduce costs, and improve customer satisfaction. Traditional forecasting methods often struggle to capture the complexities of modern retail environments, which are influenced by factors such as seasonality, promotions, competition, and external shocks. This paper explores the application of machine learning techniques to enhance demand forecasting accuracy for retail businesses. By leveraging historical sales data, external factors, and customer behavior patterns, machine learning models such as XGboost, long short-term memory (LSTM), and random forest are compared in terms of performance and adaptability to changing market conditions. The study highlights the importance of feature engineering, hyperparameter tuning, and model selection in achieving robust forecasts. Real-world case studies and experiments demonstrate the significant improvement in forecast accuracy over traditional statistical methods, offering actionable insights for inventory planning and operational efficiency. The findings underscore the transformative potential of machine learning in addressing the dynamic challenges of retail demand forecasting, paving the way for smarter decision-making and competitive advantage.

Keywords: Demand forecasting, Long Short-Term Memory, Retail Analytics, XGboost, Random Forest

1. Introduction

Accurate demand forecasting is a cornerstone of success in the retail industry, where operations are intricately influenced by fluctuating consumer behavior, seasonal trends, and dynamic market conditions. Retailers face the challenge of balancing inventory levels to minimize costs, prevent stockouts, and avoid overstocking—all while meeting the ever-evolving expectations of their customers. In such a competitive and fast-paced environment, the ability to anticipate demand with precision is not merely advantageous but essential for sustaining profitability and operational efficiency. However, traditional forecasting methods, predominantly reliant on linear models or simplistic statistical approaches, often fall short in capturing the complexity and multifaceted nature of modern retail demand patterns.

Recent advancements in machine learning (ML) have introduced transformative approaches to demand forecasting, empowering retailers to process vast datasets, uncover intricate nonlinear relationships, and respond dynamically to market shifts. Unlike traditional models, advanced ML techniques harness computational power and sophisticated algorithms to deliver superior accuracy in predictions, even in highly volatile or irregular datasets. Among these

techniques, three stand out for their distinct capabilities and proven effectiveness in retail applications:

1. **Long Short-Term Memory (LSTM)**: As a deep learning architecture, LSTM is designed to capture sequential dependencies within time-series data. This makes it particularly effective for forecasting daily sales trends, recognizing seasonality, and adapting to changing patterns over time. Its ability to remember long-term dependencies gives it a significant edge in handling complex, sequential data structures.
2. **XGBoost**: A gradient boosting algorithm known for its efficiency, scalability, and exceptional performance in structured data scenarios. XGBoost excels at modeling interactions between multiple influencing factors, such as pricing, promotions, and external events, making it a powerful tool for identifying key drivers of demand.
3. **Random Forest**: As an ensemble learning method, Random Forest combines the predictions of multiple decision trees to deliver robust and interpretable results. It is particularly valued for its ability to handle noisy data, reduce overfitting, and provide insights into feature importance, which can be instrumental in understanding the underlying factors affecting retail demand.

This study investigates the integration of LSTM, XGBoost, and Random Forest models in optimizing retail demand forecasting. By leveraging these advanced ML techniques, the research seeks to address the limitations of traditional methods, enhance the accuracy of demand predictions, and enable data-driven decisions that optimize inventory management, pricing strategies, and supply chain operations.

Furthermore, the study explores the practical implications of these techniques, examining their effectiveness in real-world retail scenarios. It provides actionable insights for retailers aiming to adopt cutting-edge forecasting models, offering a roadmap for their implementation and adaptation to specific business needs. By bridging the gap between theoretical advancements and practical applications, this research contributes to the growing body of knowledge in machine learning for retail, addressing the industry's need for agile, precise, and scalable demand forecasting solutions in an ever-changing marketplace.

1.1 Research Objective

1. Apply advanced machine learning techniques to enhance retail demand prediction.

2. Integrate external and dynamic variables (e.g., income, promotions) into forecasting models.

3. Propose a framework for effective, scalable, and accurate demand forecasting using machine learning.

2. Literature Reviews

This section provides an overview of existing research on retail demand forecasting, with a focus on the use of advanced machine learning (ML) techniques such as Long Short-Term Memory (LSTM), XGBoost, and Random Forest. The review examines the strengths and limitations of these approaches, identifies gaps in current knowledge, and highlights their application in optimizing retail operations.

Accurate demand forecasting is critical to retail operations as it directly impacts inventory management, supply chain efficiency, and customer satisfaction. Traditional forecasting methods, such as exponential smoothing and ARIMA, have been widely used in retail but often fail to capture complex, nonlinear patterns in consumer behavior, especially in the context of

modern retail dynamics (Taylor & Larrañeta, 2019). Machine learning has emerged as a robust alternative due to its ability to process large datasets and adapt to fluctuating demand patterns (Zhang et al., 2020, Areerakulkan et. al., 2024).

Machine learning models have gained significant traction in demand forecasting for their ability to identify hidden patterns in data and improve forecast accuracy. Key ML approaches applied in retail include:

- **Long Short-Term Memory (LSTM)**

LSTM networks are a type of recurrent neural network (RNN) specifically designed to handle sequential and time-series data. Studies have demonstrated their effectiveness in modeling demand patterns influenced by seasonality, holidays, and promotions. Gao et al. (2021) utilized LSTM to forecast daily sales in multi-store retail chains, achieving superior accuracy compared to traditional time-series models. Nithin et. al. (2022) introduces a hybrid Convolutional Neural Network (CNN) and LSTM model to predict stock requirements based on historical sales data. The combined architecture effectively captures spatial and temporal patterns, enhancing forecasting accuracy. Nasseri et. al. (2023) compares the performance of LSTM networks with tree-based ensemble methods in forecasting demand for perishable retail products. Utilizing over six years of historical data, the study finds that while ensemble methods like Extra Tree Regressors perform well, LSTM networks offer competitive accuracy, particularly for certain product categories.

- **XGBoost**

XGBoost, a gradient boosting algorithm, has been widely used for structured data forecasting. Its efficiency and ability to handle missing data make it ideal for retail datasets, which often contain diverse variables such as pricing, promotions, and external factors. Liu et. al. (2020) applied XGBoost to predict demand fluctuations in e-commerce platforms, highlighting its scalability and speed. Bozdogan & Alptekin (2023) presents an advanced approach to daily retail order predictions by applying multiple machine learning paradigms, including XGBoost. The study focuses on the fresh food market, highlighting the effectiveness of XGBoost in handling complex, nonlinear demand patterns inherent in perishable goods.

- **Random Forest**

Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and robustness. Its application in retail forecasting has shown promising results, particularly in feature-rich environments. Mitra et. al. (2024) used Random Forest to forecast SKU-level demand, demonstrating improved performance in handling noisy and imbalanced data. Pang (2022) utilizes the Random Forest Regressor model to forecast weekly sales for Walmart stores. The model demonstrates high accuracy, highlighting the effectiveness of Random Forest in retail demand forecasting.

- **Comparative Performance of ML Models**

Several studies have compared the performance of LSTM, XGBoost, and Random Forest in retail demand forecasting. Findings indicate:

1. **LSTM** excels in capturing sequential dependencies but requires more computational resources.
2. **XGBoost** is well-suited for structured data and provides high interpretability.
3. **Random Forest** offers robustness and is effective in feature-rich, noisy datasets. However, the choice of model depends on the specific context, data characteristics, and business requirements (Liu et al., 2021).

3. Research Methodology

This section outlines the research design and methods to be used in optimizing retail demand forecasting with advanced machine learning techniques, focusing on Long Short-Term Memory (LSTM), XGBoost, and Random Forest models. The methodology combines data collection, preprocessing, model development, evaluation, and comparative analysis to ensure comprehensive insights.

3.1 Research Design

This study employs a quantitative research design using real-world retail sales data to develop and evaluate machine learning models. The design integrates:

- Exploratory Analysis: To understand the patterns and characteristics of the data.
- Model Development: To implement and fine-tune LSTM, XGBoost, and Random Forest for demand forecasting.
- Comparative Analysis: To evaluate the performance of the models against predefined metrics.

3.2 Data Set

This section analyzes weekly sales data of 64-ounce bottled orange juice from 83 stores in the Chicago area, including numerous urban stores, across various time periods. The dataset includes three orange juice brands: Dominick's, Minute Maid, and Tropicana. The data is structured in rows, where each row represents recorded sales (in logarithmic form, logmove), along with details such as the brand, price, presence or absence of promotional advertising, and store demographic characteristics. The dataset contains a total of 28,947 rows. The details for each column in the dataset are as follows:

- **STORE**: Store number
- **BRAND**: Orange juice brand (Dominick's, Minute Maid, Tropicana)
- **WEEK**: Week number
- **LOGMOVE**: Logarithm of sales volume
- **PRICE**: Price
- **FEAT**: Product advertising (1 = Yes, 0 = No)
- **AGE60**: Proportion of the population aged 60 and above
- **EDUC**: Proportion of the population with a college degree
- **ETHNIC**: Proportion of the population that is Black or Hispanic
- **INCOME**: Logarithm of median income
- **HHLARGE**: Proportion of households with more than 5 members
- **WORKWOM**: Proportion of women working full-time
- **HVAL150**: Proportion of households with a value over $150,000
- **SSTRDIST**: Distance to the nearest store
- **SSTRVOL**: Ratio of this store’s sales to the nearest warehouse's sales
- **CPDIST5**: Average distance (in miles) to the nearest 5 supermarkets
- **CPWVOL5**: Ratio of this store’s sales to the average sales of the nearest 5 supermarkets

  (Note: The dataset can be downloaded from https://github.com/gchoi/Dataset/blob/master/oj.csv)

3.3 Data Preprocessing

- **Handling Missing Values**: Impute missing data using statistical methods or predictive models.

- **Normalization**:
  - Scale numerical data to ensure compatibility with machine learning algorithms.
- **Time-Series Preparation**:
  - For LSTM, convert data into sequences for time-series forecasting.

3.4 Machine Learning Models

Long Short-Term Memory (LSTM)

- Purpose: Capture temporal dependencies and trends in sales data.
- Implementation:
  - Develop a Seq2Seq LSTM model for multi-step forecasting.
  - Use libraries such as TensorFlow or PyTorch for model building.
- Hyperparameter Tuning: Optimize parameters like learning rate, batch size, and number of layers.

XGBoost

- Purpose: Leverage structured data to model nonlinear relationships and interactions.
- Implementation:
  - Train an XGBoost model using features derived from the dataset.
  - Use grid search or Bayesian optimization for hyperparameter tuning.

Random Forest

- Purpose: Provide robust and interpretable predictions in feature-rich environments.
- Implementation:
  - Train a Random Forest model using bootstrapped samples and feature subsets.
  - Optimize hyperparameters like the number of trees and maximum depth.

3.5 Model Evaluation

- Performance Metrics:
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - Mean Absolute Percentage Error (MAPE)
- **Validation**:
  - Split the dataset into training (80%) and test (20%) sets.
  - Use cross-validation to ensure model robustness.

3.6 Comparative Analysis

- Compare the performance of LSTM, XGBoost, and Random Forest models.
- Identify scenarios where each model excels, considering factors such as data type, complexity, and temporal dependencies.

3.7 Tools and Technologies

- **Programming Language**: Python
- **Libraries and Frameworks**:
  - Pandas and NumPy for data preprocessing.
  - Scikit-learn for Random Forest and general machine learning tasks.
  - TensorFlow or PyTorch for LSTM.
  - XGBoost for gradient boosting modeling.
- **Visualization**: Matplotlib and Seaborn for data visualization and result analysis.

4. Results

The distribution of sales (logmove) for each orange juice is illustrated using a histogram, as shown in Figure 1 below.

Figure 1: Histogram of each orange juice brand

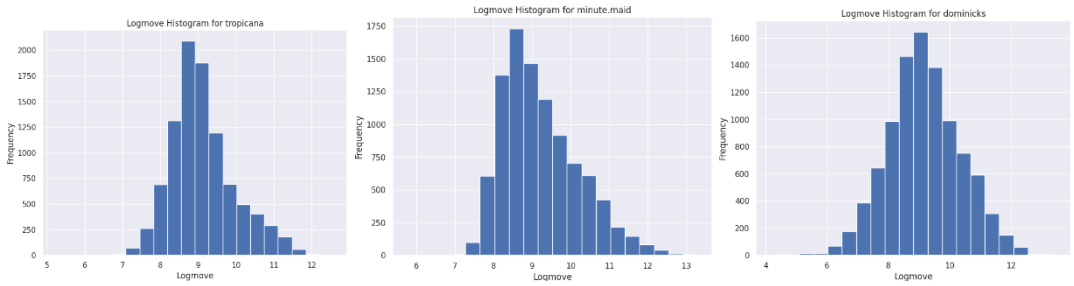
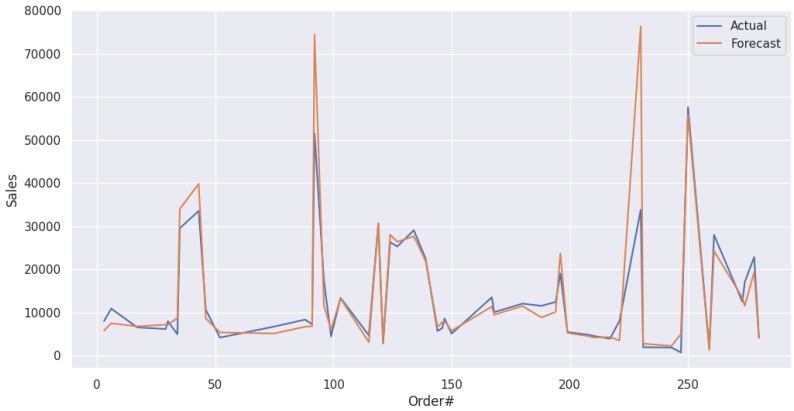


Figure 1 reveals that the log-transformed sales of Tropicana and Dominick's follow a normal distribution, while Minute Maid's sales exhibit a slight right skew. Moreover, no outliers or anomalies are detected. Based on these observations, we proceed to the next step of splitting the dataset into training (80%) and testing (20%) subsets. The models are then fitted using the Random Forest algorithm, LSTM, and XGBoost, with the results detailed in Sections 4.1 through 4.3, respectively.

1.1 Results of the Random Forest Model

The Random Forest (RF) model was trained using Python's scikit-learn library, with hyperparameter tuning conducted using the GridSearchCV function. The model was then applied to forecast sales on the testing dataset. For clarity and readability, the results of the first 275 order# are plotted and presented in Figure 2.

Figure 2: Actual vs. forecast plot



The results in Figure 2 show that the predicted values are close to the actual values. However, the accuracy of the predictions can be further improved by using the GridSearchCV function in the scikit-learn library. From the GridSearch process, the model's accuracy was improved, with the MSE decreasing from 0.165 to 0.162 and the R^2 increasing from 0.840 to 0.842. The model optimized through GridSearchCV can be utilized in the next step to analyze feature importance (as shown in figure 3) to determine which features significantly impact sales forecasting.

Figure 3: RF feature importance

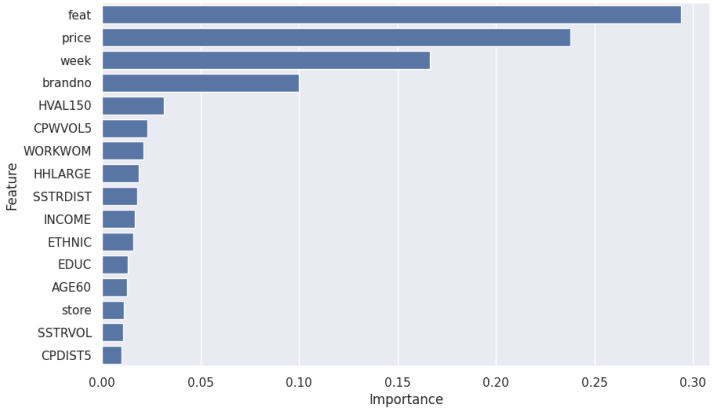


Figure 3 highlights the most influential features through feature importance analysis, allowing us to identify the key factors driving the predictions. In this case, the top four variables significantly impacting sales forecasting are **feat** (advertising), **price** (price), **week** (week number), and **brandno** (orange juice brand). High feature importance for advertising suggests that marketing efforts play a critical role in influencing sales, while the high importance of price reflects the sensitivity of customer behavior to pricing.

1.2 Results of the XGboost Model

The LSTM model was trained using Python's xgboost library, then the model applied to forecast sales on the testing dataset. The Xgboost model's accuracy was better than that of RF model, with the MSE 0.121 and the R^2 0.882. The feature importance as shown in figure 4 to determine which features significantly impact sales forecasting.

Figure 4: XGboost feature importance

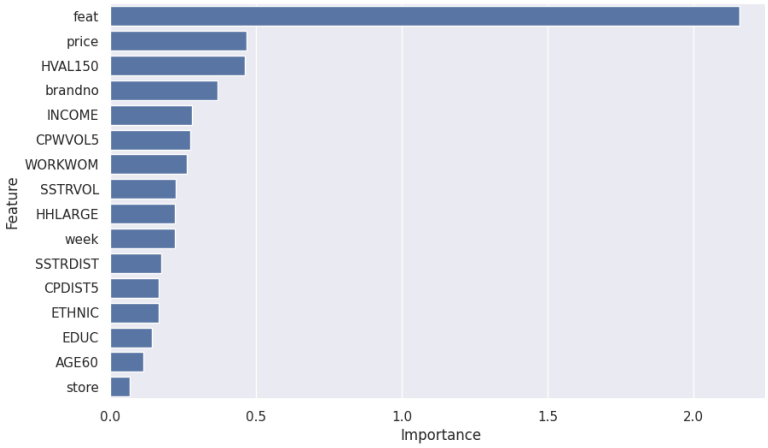


Figure 4 highlights the most influential features identified through feature importance analysis, enabling us to determine the key factors driving the predictions. In this case, the top four variables significantly impacting sales forecasting are **feat** (advertising), **price** (price), **HVAL150** (proportion of households with a value over $150,000), and **brandno** (orange juice brand). Notably, the only feature differing between XGBoost and Random Forest is **HVAL150**, while the remaining top features are consistent across both models.

1.3 Results of the LSTM Model

The LSTM model was trained using Python's tensorflow/keras library, then the model applied to forecast sales on the testing dataset. In this case, the LSTM model's accuracy was the worst, with the MSE 0.306 and the R^2 0.696. The LSTM feature importance is not directly available like in tree-based models, since LSTM is a black-box model.

5. Conclusion

This study analyzed the sales data of three orange juice brands—Tropicana, Dominick's, and Minute Maid—using advanced machine learning techniques, including Random Forest (RF), XGBoost, and Long Short-Term Memory (LSTM) models. The following summarizes the key findings and implications of the analysis:

1. **Sales Distribution**:

The log-transformed sales data for Tropicana and Dominick's exhibit a normal distribution, while Minute Maid shows a slight right skew. No outliers or anomalies were detected, enabling a clean split of the dataset into training (80%) and testing (20%) subsets for modeling.

2. **Model Performance**:
    - **Random Forest (RF)**: The RF model, tuned with GridSearchCV, demonstrated strong performance with an MSE of 0.162 and an R^2 of 0.842. Feature importance analysis identified advertising, price, week number, and brand as the key factors influencing sales, highlighting the significant impact of marketing efforts and price sensitivity.
    - **XGBoost**: The XGBoost model outperformed the RF model, achieving the best accuracy with an MSE of 0.121 and an R^2 of 0.882. The top influential features were advertising, price, HVAL150 (proportion of households with a value over $150,000), and brand. HVAL150 emerged as a unique factor not identified by the RF model, emphasizing the importance of demographic characteristics in sales prediction.
    - **LSTM**: The LSTM model exhibited the lowest accuracy among the three methods, with an MSE of 0.306 and an R^2 of 0.696. The lack of explicit feature importance due to LSTM's black-box nature limits interpretability, making it less effective for actionable insights in this context.
3. **Feature Importance**:

Across all models, advertising and price consistently emerged as the most significant features, underscoring their critical role in driving sales. The addition of HVAL150 in the XGBoost model highlights the potential influence of household demographics on sales forecasting.

4. **Implications**:
    - Marketing efforts, such as promotions and advertising, should remain a priority to enhance sales performance.
    - Pricing strategies should be carefully optimized to align with consumer sensitivity and maximize revenue.
    - Incorporating demographic data, as demonstrated by XGBoost, can improve prediction accuracy and provide deeper insights into consumer behavior.
5. **Recommendations**:

Based on the results, XGBoost is recommended as the most effective model for sales forecasting in this study. While LSTM has potential for sequential data, its performance in this

case was suboptimal, and its lack of interpretability may limit its application for actionable insights.

In conclusion, leveraging machine learning models such as Random Forest and XGBoost, coupled with feature importance analysis, provides valuable insights into sales forecasting and key drivers of performance. Future research could explore hybrid models or other deep learning architectures to improve predictive power and interpretability further.

References

Areerakulkan, N., Moryadee, C., Trakoonsanti, L., Khaengkhan, M., & Setthachotsombut, N. (2024). A new product's demand forecasting using artificial neural network. In *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS)*. https://dblp.org/db/conf/iceis/iceis2024-1.

Bozdoğan, U., & Alptekin, G. I. (2023). Demand forecasting for daily retail orders in fresh food market. *2023 4th International Informatics and Software Engineering Conference (IISEC)*, 1–5. https://doi.org/10.1109/IISEC59749.2023.10391047

Gao, X., & Wang, Y. (2021). *Automatic sales forecasting system based on LSTM network*. In *Proceedings of the 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)* (pp. 1234-1240). IEEE. https://doi.org/10.1109/IAEAC51226.2021.9443687

Liu, C. J., Huang, T. S., Ho, P. T., Huang, J. C., & Hsieh, C. T. (2020). Machine learning-based e-commerce platform repurchase customer prediction model. *PLOS ONE, 15*(12), e0243105. https://doi.org/10.1371/journal.pone.0243105

Mitra, R., Saha, P., & Tiwari, M. K. (2024). Sales forecasting of a food and beverage company using deep clustering frameworks. *International Journal of Production Research, 62*(9), 3320–3332. https://doi.org/10.1080/00207543.2023.2231098

Nasseri, M., Falatouri, T., Brandtner, P., & Darbanian, F. (2023). Applying Machine Learning in Retail Demand Prediction—A Comparison of Tree-Based Ensembles and Long Short-Term Memory-Based Deep Learning. *Applied Sciences*, *13*(19), 11112. https://doi.org/10.3390/app131911112

Nithin, S. S. J., Rajasekar, T., Jayanthy, S., Karthik, K., & Rithick, R. R. (2022). Retail demand forecasting using CNN-LSTM model. *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, 1751–1756. https://doi.org/10.1109/ICEARS53579.2022.9752283

Pang, S. (2022). Retail sales forecast based on machine learning methods. *2022 6th Annual International Conference on Data Science and Business Analytics (ICDSBA)*, 357–361. https://doi.org/10.1109/ICDSBA57203.2022.00030

Taylor, S., & Larrañeta, B. (2019). *Improving students' daily life stress forecasting using LSTM neural networks*. MIT Media Lab. https://www.media.mit.edu/publications/improving-students-daily-life-stress-forecasting-using-lstm-neural-networks/

Zhang, G., Ren, T., & Yang, Y. (2020). *A new unified deep learning approach with decomposition-reconstruction-ensemble framework for time series forecasting*. arXiv. https://arxiv.org/abs/2002.09695