# Analysis and Comparison of Fresh Sales Forecasting Models Based on FreshRetailNet-50K Dataset

Qiong Wu *

Software College, Northeastern University, Shenyang, 110000, China

* Corresponding Author Email: eieieikoi@outlook.com

**Abstract.** In the context of the increasingly intertwined digital economy and traditional retail businesses, the role of data-driven decision-making has become a vital strategy to improve the efficiency of the fresh produce chain. In this paper, the authors systematically analyse the differences between the following three popular sales prediction models: Prophet, Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory Model (LSTM), based on the openly available FreshRetailNet-50K dataset. The dataset covers the multidimensional feature set information, as well as the external influencing factors such as weather information, promotional activities, and holiday information. From the experiment results, the authors find that the prediction accuracy of the XGBoost model ranks the highest among the three sales prediction methods, with the lowest Root Mean Square Error (RMSE) at 0.714 and the highest R-Square ($R^2$) at 0.816 compared to the Prophet and LSTM counterparts. Moreover, the authors apply the SHapley Additive exPlanations (SHAP) method to explain the interpretability of the above results and find that the time series-based features like “Lag Sale” and “Rolling Average Sale” play the most important role in determining the prediction results and demonstrate the "inertia effect" of the sales in the very short term.

**Keywords:** Sales Forecasting, XGBoost, LSTM, Prophet, SHAP Explainability.

## 1. Introduction

In the context of the deeper integration between the digital economy and traditional retailing, the application of data analytics has become a vital drive for businesses to improve efficiency and competitiveness. Especially in the fresh product retail industry, the vegetables' shorter shelf life and fluctuation patterns create specific difficulties in logistics management (Wang and Zhang, 2021). In contrast to traditional products, the shorter shelf life of vegetables indicates that surplus stock immediately leads to lost investment when not sold in time. On the other hand, understocking indicates the loss of sales and customer satisfaction(Accorsi and Manzini, 2021). This has made the prediction of future sales information for optimization and corresponding marketing activities a burning problem to be settled in the fresh retail business.

Academic and industrial circles have thoroughly researched the area of sales forecast challenges. Methods such as the Prophet model can analyse trends and seasonality components in the time series. They perform very well when the cyclical pattern has been repeating uniformly(Kumar and Singh, 2023). This can be less effective when the model has to incorporate many external factors. For instance, the factors can include promotions and weather (Taylor and Letham, 2018).

In the face of the above challenges, a new set of solutions has been developed based on the principles of machine learning. One such solution is the XGBoost model. The effectiveness of the XGBoost model relies on the effective determination of the nonlinear relationships between the factors involved. This comes as a fundamental advantage when applied to the prediction of information characterized as diverse (Bentéjac et al, 2021). For example, the effectiveness of tree-based models in the prediction of information characterized as diverse has been demonstrated in the research carried out by Chen et al. According to the findings of the aforementioned research, tree-based models performed significantly better than linear models and traditional time series methods when the sales information from the context of e-commerce and diverse promotional information were involved (Chen et al, 2021;Kumar and Singh, 2023).In view of the above-mentioned concepts and based on the fact that the context under review involves the retailing of fresh produce and other

factors as diverse as those involved in the context of the aforementioned publication, the following hypothetical assumption can be made.

The emergence of deep learning has made time series prediction accessible through options such as LSTM, whose distinct gating systems have huge potential for storing long sequences and extracting long-term relationship patterns (Siami-Namini et al., 2021). For example, comparison analyses carried out by Zhang et al. confirmed that the pattern recognition ability of LSTM outperforms traditional machine learning techniques when dealing with long sequences of financial information (Zhang et al., 2021). This indicates that LSTM can extract more hidden information when dealing with the FreshRetailNet-50K database, since the database consists of the total sales conducted at the end of each month.

However, although the prediction accuracy can be improved through the use of machine learning and deep learning approaches, the "black box problem" can remain. In the business world, only achieving the accurate prediction number isn't sufficient; instead, "why the prediction was made" has the equal importance as the prediction outcome. Explainable Artificial Intelligence provides the correct solution to the above-mentioned challenge. Among the techniques of Explainable Artificial Intelligence, SHAP provides the attribution analysis for the prediction outcome based on the contribution of each feature(Ma et al., 2022).

In other words, although previous studies have investigated the applications of each model separately, a problem remaining unresolved in the field of retail analysis can now be addressed because the proposed thesis systematically evaluates the performance of the Prophet model compared to XGBoost and LSTM techniques in the context of the FreshRetailNet-50K dataset and uses the SHAP technique to interpret the attribution of the best-performing model..

## 2. Research Design and Methods

The dataset used for this case study is the publicly available FreshRetailNet-50K. This dataset provides information across 18 cities, 898 stores, and 863 fresh product SKUs. The resolution of this dataset goes up to the per-hour level. This dataset provides a thorough set of multidimensional features like promotions, weather, holidays, and stockouts. Such a robust set of features makes this dataset a great foundation for the case study.

### 2.1. Data Preprocessing and Feature Engineering

In the preprocessing of the data, the timestamps (dt) were processed for the extraction of time-related features representing the cyclical pattern of sales. In order to allow the model to handle the categorical nature of the features, the non-numeric columns store_id and product_id were label-encoded. In view of the natural dependency of the sales-related data following the time-based pattern, two sets of time-related features were derived: lag-based features like lag_1 for the sales made in the previous time unit representing the sales momentum and rolling statistical features like rolling_24 representing the rolling 24-hour average sales.

Starting from the creation of a candidate pool featuring more than ten features, the current research used the feature importance assessment tool provided in the XGBoost algorithm for the pre-filtering process. In the end, the eight most important factors influencing the sales forecast were chosen for the modeling steps.

The final dataset was divided into training set (60 days), validation set (15 days), and the test set (15 days).

### 2.2. Model Selection and Construction

In order to conduct a comparison among the performances of the different models used in the forecast of the sale of fresh produce, the following three models were chosen: Prophet Model, the XGBoost Model, and the LSTM Model.

Prophet has been used as the baseline model to compare the other two. A seasonality component of both the daily and weekly cycles has been enabled at the initial stage to increase the capacity of the model to predict the changing pattern. Also, the flexibility for the change in the trend given in the changepoint_prior_scale has been fixed at 0.05.

The XGBoost algorithm has strong tableau data understanding capabilities and can detect complex nonlinear interactions between features. In fact, the type of data used in this experiment represents general tableau data. In general, tableau data can consist of many forms of features. For instance, the featured type can represent time series data. In this experiment, the important parameters of the XGBoost algorithm used were set based on grid search and cross-validation techniques. These parameters were set as follows: the learning_rate factor to 0.05 in order to promote proper model convergence. The max_depth factor was set to 9. In order to add a touch of randomness during each iteration, the subsample and colsample_bytree factors were set to 0.8 and 0.7. The maximum training iterations during the experiment were set to 1000.

The “gating mechanism” of LSTMs allows them to recognize long-range patterns in the sequence. Hence, the inclusion of LSTMs in the proposed solution to analyze whether LSTMs can reveal distinct advantages over traditional solutions in understanding the long-term changes in the sales pattern. In order to prevent the overcomplexity of the network and at the same time ensure the network has sufficient capacity to perform the training job, the network uses a two-layer stacked structure. Each layer consists of 128 hidden units. In order to prevent overfitting, the network uses dropout rates of 0.2 between the hidden layers. During training, the network uses the “Adam” optimizer. The “learning rate” during training has been set between 0.001 and 0.005. While training the model, the “ReduceLROnPlateau” “learning rate” schedule has been used. In addition, the “early stopping” strategy has been used during the training process. The “training batch” has been set to 256.

### 2.3. Interpretability Analysis Methods

Though the prediction accuracy is important, the explanations for the predictions made by the model are more important. Hence, the goals of this research include the interpretation of the best-performing model. SHAP uses the Shapley value from the mathematical field of game theory. This method can "attribute" the prediction outcome to each feature of the input. This can explain the contribution of each feature to the prediction outcome. As mentioned earlier, the advantages of the SHAP method include "model independence" and "result consistency." In fact, the SHAP method can explain complex "black box" patterns. This includes the XGBoost pattern.

In order to fairly select the best-performing type of regression model among the three based on the predictability of the actual values and the corresponding predictions made by the regression model, this study uses four general regression comparison measures: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination ($R^2$), and Weighted Average Percentage Error (WAPE).

## 3. Results and Analysis

### 3.1. Model Performance Comparison

After the entire process of preprocessing the data and training the models, the prediction accuracy of the Prophet model, the XGBoost model, and the LSTM model has been tested. The specific test results are shown below in Table 1.

**Table 1:** Performance evaluation of different models on the test set.

| Model | RMSE | MAE | $R^2$ | WAPE |
|---|---|---|---|---|
| Prophet | 2.287 | 1.048 | -0.752 | 93.60 |
| XGBoost | 0.714 | 0.372 | 0.816 | 33.17 |
| LSTM | 1.266 | 0.608 | 0.706 | 54.24 |

As can be seen in Table 1 below, the best-performing model in all four measures was the XGBoost. Its Root Mean Squared Error of 0.714 reflects a percentage reduction of 43.6% relative to the LSTM model's 1.266 and a percentage decrease of 68.8% compared to the baseline Prophet model of 2.287. Turning to the measures of fit of the model to the observed values, the R-squared value of the XGBoost at 0.816 reflects a marked increase compared to the 0.706 recorded in the case of the LSTM. Moreover, the negative R-squared measures of the Prophet Model reveal that the prognostication prowess of the model lagged even the mean model. In fact, the Weighted Average Absolute Percentage Error measures a marked disparity in the percentage errors among the different models. While the percentage errors of the other two models were 54.24% and 93.60% respectively, that for the XGBoost Model stands at 33.17% only.

This finding strongly confirms the fundamental research hypothesis: In cases like fresh produce retailing, where complex factors of multiple dimensions are common, ML techniques proficient in the effective analysis of tabular structure and nonlinear interactions between diverse factors perform better than traditional time series analysis techniques as well as some DL techniques. However, due to the strong nonlinear analysis ability of XGBoost, the model succeeded in extracting the complex mapping between the high-dimensional features and succeeded in making the most accurate predictions. Even though the LSTM model has strong ability to analyse long-term dependencies in the tabular structure, it failed to perform better than the XGBoost model. This could be due to the fact that for the specific one-hour prediction used in this experiment, the most dominant factors influencing sales were observed to exist in shorter time frames (such as today's weather patterns and the previous promotion activities)—which exactly XGBoost excels at. In contrast, the Prophet model performed poorer than other techniques due to its intrinsic limitations in dynamically incorporating multiple regression factors.

### 3.2. XGBoost Model Interpretability Analysis

As the XGBoost model yielded the best prediction results, the current study uses the SHAP technique to analyse the factors that have the most influence on the sales of fresh produce.

Figure 1 provides a useful comparison of the relative importance of each feature. Figure 1 illustrates that the "lag_1" (previous hour sales) and "rolling_24" (24-hour rolling average) features are the most important factors. This indicates that sales of fresh produce are strongly autocorrelated. Another factor that comes close to the top feature ranks is "weekday." This can identify whether the days of the week or their factors play a role in influencing the sales of fresh produce. This indicates that the days of the week can influence the sales of the products. Another important factor that ranks high is "discount" and "avg_temperature." This has been validated based on the fact that weather factors play important nonlinear roles in influencing retail sales based on the previous studies mentioned above (Huang et al., 2023).

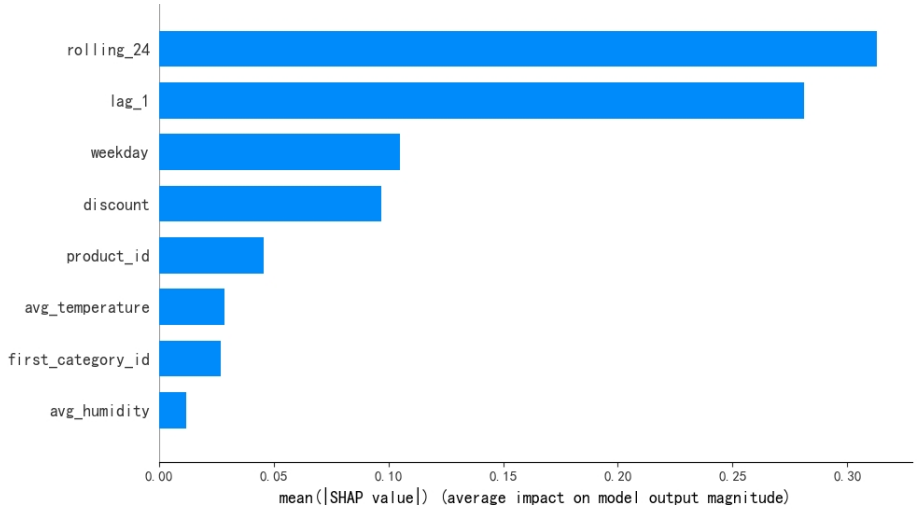


**Figure 1:** SHAP feature importance ranking chart.

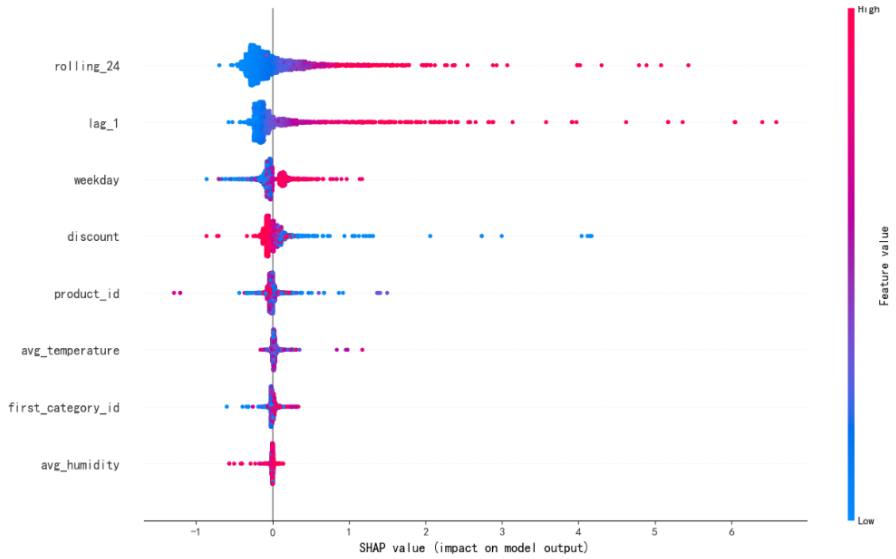


**Figure 2:** SHAP feature impact distribution map.

To further investigate how specific features influence prediction outcomes, this study generated SHAP feature impact plots.

From the above graph (Figure 2), the following can also be observed: the effect of the values of the features on the predictions. Each dot represents a test case. The colour of the dots represents the magnitude of the feature values. The positioning of the dots on the X-axis represents the effect the feature has on the prediction (positive or negative). For example, in the case of rolling_24, the effect of low rolling average sales (blue dots) is nearly solely in the negative region of the X-axis. This has a profoundly negative effect. In contrast, the effect of high rolling average sales (red dots) has a profoundly positive effect. This serves as the perfect example of the "Matthew Effect" observed in sales: bestsellers become bestsellers again, and the opposite happens to under-performing items (Rigney, 2010).

To more precisely and intuitively illustrate the functional relationship between the core feature "rolling_24" and its impact, this paper further plots its SHAP value map (Figure 3).

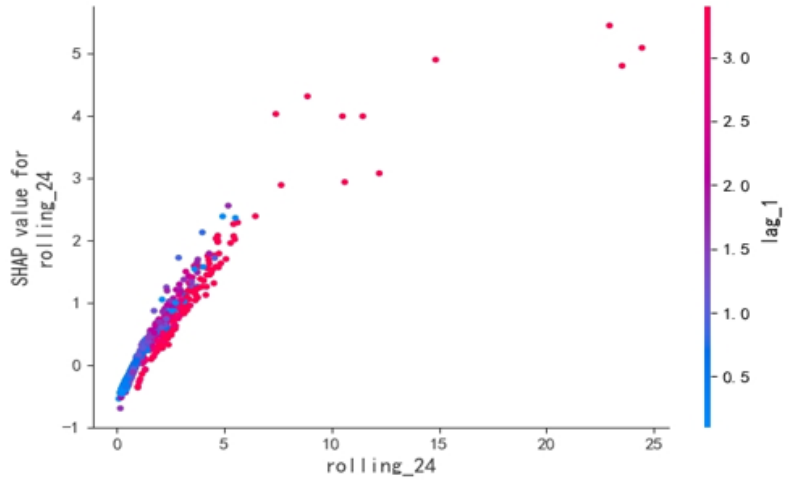


**Figure 3:** SHAP dependency map for rolling_24.

As can be seen from Figure 3 above, the graph clearly shows the linear positive correlation trend between the variables, functionally proving again that the more positive the sales momentum in the latest sales, the more positive the prediction outcomes for the sales. Moreover, the colour used in each point in the graph shows the value of the lag_1 feature that has the strongest interaction. In the high rolling_24 category, the SHAP values shown in the graph have predominantly red-coloured points; this indicates that when the 24-hour rolling sales and 1-hour sales values are high, their effects complement each other.

In other words, this paper not only proves the effectiveness of the XGBoost model but also uses the SHAP technique to “break open” the “black box” solution. This verification shows the important role of sales records as the focal point of the prediction. This output can directly inform businesses how to create retail inventory marketing solutions.

## 4. Conclusion

Based on the FreshRetailNet-50K dataset, this paper systematically investigates sales forecasting in fresh retail scenarios, comparing three representative models: Prophet, XGBoost, and LSTM. Experimental results indicate that the XGBoost model demonstrates superior performance in both prediction accuracy and stability, exhibiting the lowest RMSE and highest $R^2$. It effectively captures nonlinear relationships and interactions among multidimensional features. In contrast, while the LSTM model possesses temporal dependency modeling capabilities, its advantages are not pronounced in the hourly sales data analysed here. The Prophet model shows limitations when handling complex external factors. SHAP interpretability analysis further reveals that "sales from the previous hour (lag_1)" and "24-hour rolling average sales (rolling_24)" are the core variables influencing prediction outcomes, reflecting the short-term inertia characteristics of sales. Additionally, external factors such as promotional discounts, weather changes, and holidays significantly influence sales fluctuations. This finding not only validates the strengths of machine learning models in analysing high-dimensional retail data but also provides data support and theoretical references for enterprises in dynamic inventory management and precision marketing strategy formulation.

Overall, this study reveals the applicability and differences of various forecasting models in the fresh food retail sector from both methodological and practical perspectives, establishing a replicable experimental framework for future research. Future work may incorporate additional heterogeneous data sources and advanced deep learning architectures to develop sales forecasting models with higher accuracy and stronger generalization capabilities.

## References


[1] Accorsi, R., & Manzini, R. (2021). A review of data-driven approaches for fresh food supply chain management. Trends in Food Science & Technology, 108, 1–15.

[2] Bentéjac, C., Csörgö, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937–1967.

[3] Chen, Y., Li, F., & Wang, J. (2021). A comparative analysis of machine learning models for e-commerce sales forecasting with promotion information. Journal of Retailing and Consumer Services, 62, 102643.

[4] Huang, X., Zhang, Q., & Li, J. (2023). Weather and promotion effects on retail sales: Evidence from multivariate forecasting. Decision Support Systems, 164, 113826.

[5] Kumar, A., & Singh, R. (2023). Retail demand forecasting using Prophet and ARIMA models: A comparative study. In Proceedings of the 4th International Conference on Advances in Computing and Data Sciences (ICACDS).

[6] Ma, X., Liu, Z., & Sun, Y. (2022). Interpretable freight volume forecasting with spatio-temporal graph neural networks and SHAP. IEEE Transactions on Intelligent Transportation Systems, 23, 19579–19589.

[7] Rigney, D. (2010). The Matthew Effect: How advantage begets further advantage. Columbia University Press.

[8] Siami-Namini, S., & Namin, A. S. (2021). A comparative analysis of forecasting financial time series using ARIMA, LSTM, and GRU. Journal of Risk and Financial Management, 14(12), 615.

[9] Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37–45.

[10] Wang, J., & Zhang, H. (2021). Integration of digital economy and physical retail: Opportunities and challenges for data-driven decision making in fresh produce supply chains. Journal of Business Research, 130, 456–466.

[11] Zhang, L., Wang, Y., & Gao, S. (2021). A comparative study of LSTM, GRU, and transformer models for stock price prediction. Journal of Risk and Financial Management, 14, 365.