EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Research on Fresh Produce Sales Prediction Algorithm for Store Based on Multidimensional Time Series Data Analysis

Li Zhiyu[1], Yu Wei[1,*], Zhu Wenwei[1], Wan Haojie[1], Peng Jingjing[1], Wang Hui[2]

[1] School of Computer Science, Wuhan University, Wuhan, Hubei Province, China
[2] Hubei Key Laboratory of Big Data in Science and Technology (Wuhan Library of Chinese Academy of Sciences), Wuhan, Hubei Province, China

Abstract

INTRODUCTION: Fresh produce is a daily necessity; however, offline stores often rely on personal experience for purchase, which is highly subjective and may result in inaccurate estimation of purchase quantities. This can lead to produce wastage and subsequently impact the profitability of business. This paper proposes the Fresh Produce Sales Prediction Framework, which can predict fresh produce sales by analyzing multidimensional time series data that influence sales. This model aims to provide guidance for fresh produce purchase in offline stores.
OBJECTIVES: The purpose of this study is to predict fresh produce sales by analyzing multidimensional time series data that influence sales. This aims to provide a basis for fresh produce purchase in stores, reduce produce wastage, and enhance business profitability.
METHODS: This paper proposes the Fresh Produce Sales Prediction Framework by analyzing multidimensional time series data that affect store sales of fresh produce. An essential component of this model is the ARIMA-LSTM Combined Prediction Model. In this study, the Weighted Reciprocal of Errors Averaging Method is selected as the weight determination method for the ARIMA-LSTM Combined Prediction Model.
RESULTS: In this paper, the ARIMA-LSTM combined model is used for prediction in two scenarios: when the single-model prediction accuracy is superior and when it is inferior. Experimental results indicate that in the case of lower accuracy in single-model prediction, the ARIMA-LSTM Combined Prediction Model outperforms, improving prediction accuracy by 3.86% as measured by MAPE. Comparative experiments are conducted between the Fresh Produce Sales Prediction Framework proposed in this paper and time series prediction framework Prophet, traditional LSTM model, and ARIMA model. The experimental results indicate that the proposed model outperforms the others.
CONCLUSION: The Fresh Produce Sales Prediction Framework proposed in this paper is based on multidimensional time series data to predict fresh produce sales in stores. This model can accurately predict the sales of fresh produce, providing purchase guidance for fresh produce stores, reducing fresh produce wastage caused by subjective purchasing factors, and increase business profits.

Keywords: Fresh produce sales prediction, Multidimensional time series data, Combined prediction model, LSTM

Received on 15 September 2023, accepted on 19 October 2023, published on 20 October 2023

doi: 10.4108/eetsis.3844

1. Introduction

Fresh produce is a daily necessity characterized by seasonality, concentrated availability, and a short shelf life [1]. Offline fresh produce stores often over-purchase in order to meet customer demands, leading to the spoilage of some items, resource wastage, and increased operational costs. Currently, offline stores primarily rely on historical experience and current sales data to estimate future sales for the upcoming days. They then combine this with stock to determine purchase quantities and place purchase orders. These orders are fulfilled by distribution warehouses, and the fresh produce needs to be sold quickly once it arrives at the store. The longer items stay in the store, the lower their

* Corresponding author. Email: yuwei@whu.edu.cn

quality and profitability. Prolonged retention beyond the freshness period results in waste. Manual purchase is subjective and random, often deviating significantly from actual demand. Factors such as store size, location, and other variables vary from one store to another, making it challenging to establish accurate empirical rules. This creates inconvenience for stores in purchasing fresh produce.

Prediction refers to the quantified estimation and projection of potential future events using existing knowledge and methods. It can scientifically and objectively reflect the inherent characteristics and developmental rules of things, thus providing guidance when dealing with complex and unknown scenarios [2]. Prediction methods can be broadly categorized into traditional prediction and intelligent prediction [3]. Traditional prediction methods include techniques like regression analysis and ARIMA models. Intelligent prediction methods primarily rely on machine learning, such as BP neural network models and LSTM neural network models [4]. Traditional methods perform well in handling short-term linear data but may underperform when dealing with medium to long-term nonlinear data. In contrast, intelligent methods excel in addressing nonlinear data but may face challenges like overfitting when dealing with smaller sample sizes. Therefore, the choice of an appropriate prediction method depends on the specific data characteristics and the type of problem being studied, with the goal of improving the accuracy of prediction results.

This paper conducts research on fresh produce sales prediction methods based on data from the backend database of a supermarket's fresh produce sales system in a city in Hubei Province, China. This paper proposes the Fresh Produce Sales Prediction Framework (FPSPF), which analyzes multidimensional time series data that affect fresh produce sales to predict the sales of fresh produce. This paper analyzes the advantages and limitations of ARIMA and LSTM models and, considering the characteristics of fresh produce sales, proposes the ARIMA-LSTM Combined Prediction Model based on the ARIMA time series model and the LSTM neural network model. This combined model is a crucial component of the Fresh Produce Sales Prediction Framework. Therefore, this paper validates its feasibility in scenarios where the single-model prediction performs both superior and inferior. Experimental results indicate that the combined model performs better when the single-model prediction has lower accuracy.

2. Related Work

Novianti Trisita et al. [5] utilized historical demand data for salt in Indonesia and employed an ARIMA model to predict the demand for exported salt in Indonesia and provided the optimized model to managers in the salt manufacturing industry. This effectively reduced the impact of the bullwhip effect on the supply chain. Longzhou Chen [6] analyzed historical price data for pork in the Shanghai wholesale market using the ARIMA model. Through residual analysis, they demonstrated that the ARIMA model had a good fit for the data. Zhao Xin [7] utilized 20 years of monthly container throughput data in Tianjin, China, and employed the ARIMA model to predict port container throughput. This provided guidance for government departments and shipping companies in port planning. Arunraj etc. [8] developed a seasonal autoregressive integrated moving average with external variables (SARIMAX) model to forecast daily sales of a perishable food. Reshid [9] made GDP, GDP growth rate, and inflation rate predictions for Ethiopia by constructing ARIMA models. Kuhe and Obed [10] used the ARIMA model to predict the hepatitis B infection rate among blood donors in a specific location in Nigeria and achieved favorable prediction results. Goyal Megha et al. [11] employed an ARIMA model to predict the export of agricultural products in India, revealing a high level of prediction accuracy with errors ranging between 2% and 4%. Ahmad T E et al. [12], in their study of fisheries prediction, found that the ARIMA model outperformed regression models. Posch etc. [13] proposed a forecasting approach that is solely based on the data retrieved from point-of-sale systems and allows for a straightforward human interpretation, they proposed two generalized additive models for predicting future sales.

It can be observed that in specific problems, traditional prediction methods often yield satisfactory results when time series data exhibit clear linear characteristics. However, these methods have noticeable limitations. When time series data also possess nonlinear features, relying solely on traditional prediction methods may result in subpar predictions. To address this situation, many scholars have introduced intelligent prediction methods, among which neural network models are widely used in practical applications. These models are particularly suited for handling data with nonlinear features.

Hossain M S and Mahmood H [14] employed LSTM models along with Recurrent Neural Networks, Generalized Regression Neural Networks, and Extreme Learning Machines for photovoltaic power generation prediction. They compared the prediction results and validated the superiority of the LSTM model. Siami-namini S et al. [15] constructed and predicted using both LSTM and ARIMA models. In comparison to the ARIMA model, the LSTM model significantly improved prediction accuracy. These results indicate that in specific problems, deep learning algorithms such as LSTM outperform traditional algorithms like ARIMA. Lee et al. [16] developed and compared the performance of three sales forecasting models based on Logistic Regression (LR), Moving Average (MA), and Back-Propagation Neural Network (BPNN), among other prediction methods, for forecasting fresh food sales in a point of sale (POS) database for convenience stores.

Absar Nurul et al. [17] utilized recovery and confirmed COVID-19 data from Bangladesh and conducted experiments to demonstrate that the LSTM model outperforms contemporary techniques in accurately predicting the recovery and confirmed case numbers of the virus. Shahani Niaz Muhammad et al. [18] modeled the drilling rate index using Long Short-Term Memory networks, Simple Recurrent Neural Networks, and Random Forest algorithms. By comparing the performance metrics of various methods, they

proved that LSTM's predictive output was superior. Pei Ying et al. [19] proposed a multi-dimensional analysis-based LSTM prediction algorithm when studying trends in direct current measurement systems. They conducted experiments using actual operational data and demonstrated that this algorithm outperformed traditional time series models in predicting drive current. Wang etc. [20] combined the perspective of People-Goods-Scene and the push-pull theory and proposes a two-stage method for forecasting sales volumes using structural equation models and artificial neural networks.

J.M. Bates and C.W.J. Geanger [21] first introduced the concept of combined prediction models in 1969. Chandriah etc. [22] proposed Recurrent Neural Networks/ Long Short-Term Memory (RNN / LSTM) with modified Adam optimizer to predict the demand for spare parts. In this LSTM, weight vectors are generated respectively. These weights are optimized using the Modified-Adam algorithm. Guo Yan et al. [23] proposed a combination model consisting of Long Short-Term Memory networks and Chebyshev polynomials. They applied this model to precipitation forecasting in a certain city and found that its outperformed LSTM networks with smaller prediction errors. Luo [24] proposed the xDeepFM-LSTM combined forecasting model for the characteristics of sales data of apparel retail enterprises. This paper first used the Extreme Deep Factorization Machine (xDeepFM) model to explore the correlation between the sales influencing features as much as possible, and then modeled the sales prediction. Next, this paper used the Long Short-Term Memory (LSTM) model for residual correction to improve the accuracy of the prediction model. Wang etc. [25] proposed a sales prediction model, M-GNA-XGBOOST, using the time-series prediction that ensures the effective prediction of sales about each product in a short time on online stores based on the sales data in the previous term or month or year. The proposed M-GNA-XGBOOST model serves as an adaptive prediction model, for which the instant factors and the sales data of the previous period are the input, and the optimal computation is based on the proposed methodology. Ahmed etc. [26] presented a novel DL approach for time series prediction using a combination of poly-linear regression with Long Short-Term Memory (LSTM) and data augmentation. It is consequently named Polylinear Regression with Augmented Long Short-Term Memory Neural Network (PLR-ALSTM-NN). The proposed DL model can be exploited to predict the future financial markets more accurately than existing state-of-the-art neural networks and machine learning tools.

It is evident that many scholars have applied combined prediction to practical problems, where combined prediction often yields better results than single-model predictions. However, the predictive performance of combined models may not always surpass that of some well-performing individual models. In specific practical scenarios, a comparative analysis is necessary to determine the optimal solution.

3. Model

3.1. ARIMA

Autoregressive Integrated Moving Average (ARIMA) is a time series analysis and prediction model. The primary objective of this model is to model time series data for the purpose of making future predictions. The ARIMA model typically consists of three main components: the Autoregressive (AR) component, the Integrated (I) component, and the Moving Average (MA) component.

The Autoregressive (AR) component represents the relationship between the observations at the current time point and past observations. For a time series $\{x_t : t \in T\}$, if it satisfies:

$$x_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + \mu_t . \quad (1)$$

In that case, it can be referred to as an AR(p) model, where $\{a_0, a_1, \cdots, a_p\}$ are the autoregressive coefficients, and μ_t is the white noise sequence coefficient.

The Integrated (I) component is used to handle non-stationary time series data and transform it into a stationary time series. The Integrated operation in an ARIMA model can be expressed as:

$$y_t = (1 - B)^d x_t . \quad (2)$$

Where B is the lag operator, and d represents the number of integrate. If the time series is stationary, d is 0, indicating that no integrate is required.

The Moving Average (MA) component represents the relationship between the current observation at time t and past white noise error terms. For a time series $\{x_t : t \in T\}$, if it satisfies the following conditions:

$$x_t = \mu_t + b_1 \mu_{t-1} + b_2 \mu_{t-2} + \cdots + b_q \mu_{t-q} . \quad (3)$$

In that case, it can be referred to as an MA(q) model, where $\{b_1, b_2, \cdots, b_q\}$ are the moving average coefficients, and μ_t is the white noise sequence coefficient.

3.2. LSTM

The Long Short-Term Memory (LSTM) is an important variant of recurrent neural networks (RNNs) in the field of deep learning, used for modeling and predicting sequential data. The core structure of an LSTM includes three key gate control components: Input Gate, Forget Gate, and Output Gate. These gate units help the LSTM network effectively capture and retain important information from the input sequence while suppressing unnecessary information. The core structure is illustrated in Figure 1:

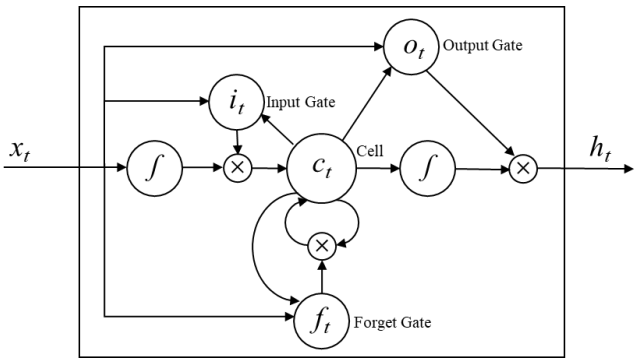

Figure 1. LSTM's Core Structure

The Input Gate is responsible for deciding which information should be stored in the Long-Term Memory (Cell State). It uses a Sigmoid activation function to control the weights of input information and uses the hyperbolic tangent function to map the input information to values within a specific range for storage in the Long-Term Memory.

The Forget Gate determines whether the information in the previous time step's Long-Term Memory should be retained or forgotten. It uses a Sigmoid activation function to determine the degree of forgetting, where 0 represents complete forgetting, and 1 represents complete retention.

The Output Gate decides which information from the Long-Term Memory should be passed to the current time step's Short-Term Memory (Hidden State). Similar to the Input Gate, it uses Sigmoid and hyperbolic tangent functions to control the output.

The variables in Figure 1 are as follows:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f), \quad (4)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i), \quad (5)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o), \quad (6)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c), \quad (7)$$

$$h_t = o_t + \sigma_h(c_t). \quad (8)$$

In which, W, U, and b are matrices learned during training, serving as elements for computation within the network. σ represents the Sigmoid function.

In summary, the working principle of LSTM can be summarized as follows: Input sequences are fed into the LSTM network. The Input Gate determines which information should be stored in the Long-Term Memory, while the Forget Gate decides which information should be forgotten. The Long-Term Memory is updated and maintained to adapt to changes in the input sequence. The Output Gate controls how information from the Long-Term Memory is passed to the current time step's Short-Term Memory. Finally, the output from the Short-Term Memory is used for prediction or the next step in sequence modeling tasks.

3.3. Fresh Produce Sales Prediction Framework

This paper introduces the Fresh Produce Sales Prediction Framework, as illustrated in Figure 2. This model involves data collection for fresh sale, text processing, data cleaning, and feature extraction. It also incorporates a data quality optimization method for few samples in real-world scenarios. Additionally, it leverages the ARIMA-LSTM Combined Prediction Model to further enhance predictive accuracy.

The specific workflow includes:

(i) This Factor analysis of fresh sales
Through on-site investigations and interviews, this paper theoretically analyzes 17 potential sales influencing factors. Data is collected, processed, and cleaned to transform it into a format suitable for analysis and experimentation. Finally, 12 effective influencing factors and appropriate text processing methods are selected through Spearman correlation analysis and time series analysis.

(ii) Data quality optimization method for few samples.
Using time series similarity analysis based on Dynamic Time Warping (DTW) algorithm, the similarity of sales between different stores is calculated to determine whether the sales patterns of a certain type of fresh produce are similar across different stores. If they are similar, the sales data of various stores are concatenated before model training.

(iii) ARIMA-LSTM Combined Prediction Model.

(a) Construction and prediction of LSTM model. Analysis and selection of features are performed for the LSTM model. The network design of the LSTM model is determined, including the number of neural network layers and deep learning hyperparameters. The trained LSTM model is used to predict daily fresh sales, and prediction errors are evaluated. If the LSTM model achieves satisfactory accuracy, i.e., MAPE $\leqslant$ 25% or RMSE $\leqslant$ 50, it is selected as the final model. If the accuracy falls short, i.e., MAPE > 25% and RMSE > 50, further steps are taken to improve accuracy.

(b) Construction and Prediction of ARIMA model. The input for the ARIMA model does not requires additional preprocessing. Only the daily sales data obtained during the LSTM preprocessing is used. The data is subjected to stationarity transformation, and model parameters are determined using the BIC criterion. The model is trained to obtain predictions of daily sales for fresh produces.

(c) Construction and Prediction of the ARIMA-LSTM Combined Prediction Model. In this paper, the weights of the combined model are determined using the Weighted Reciprocal of Errors Averaging Method, and subsequently, the ARIMA-LSTM Combined Prediction Model is constructed. If this model outperforms the LSTM model in terms of prediction accuracy, the ARIMA-LSTM Combined Prediction Model is chosen as the final model; otherwise, the LSTM model is selected as the final model.

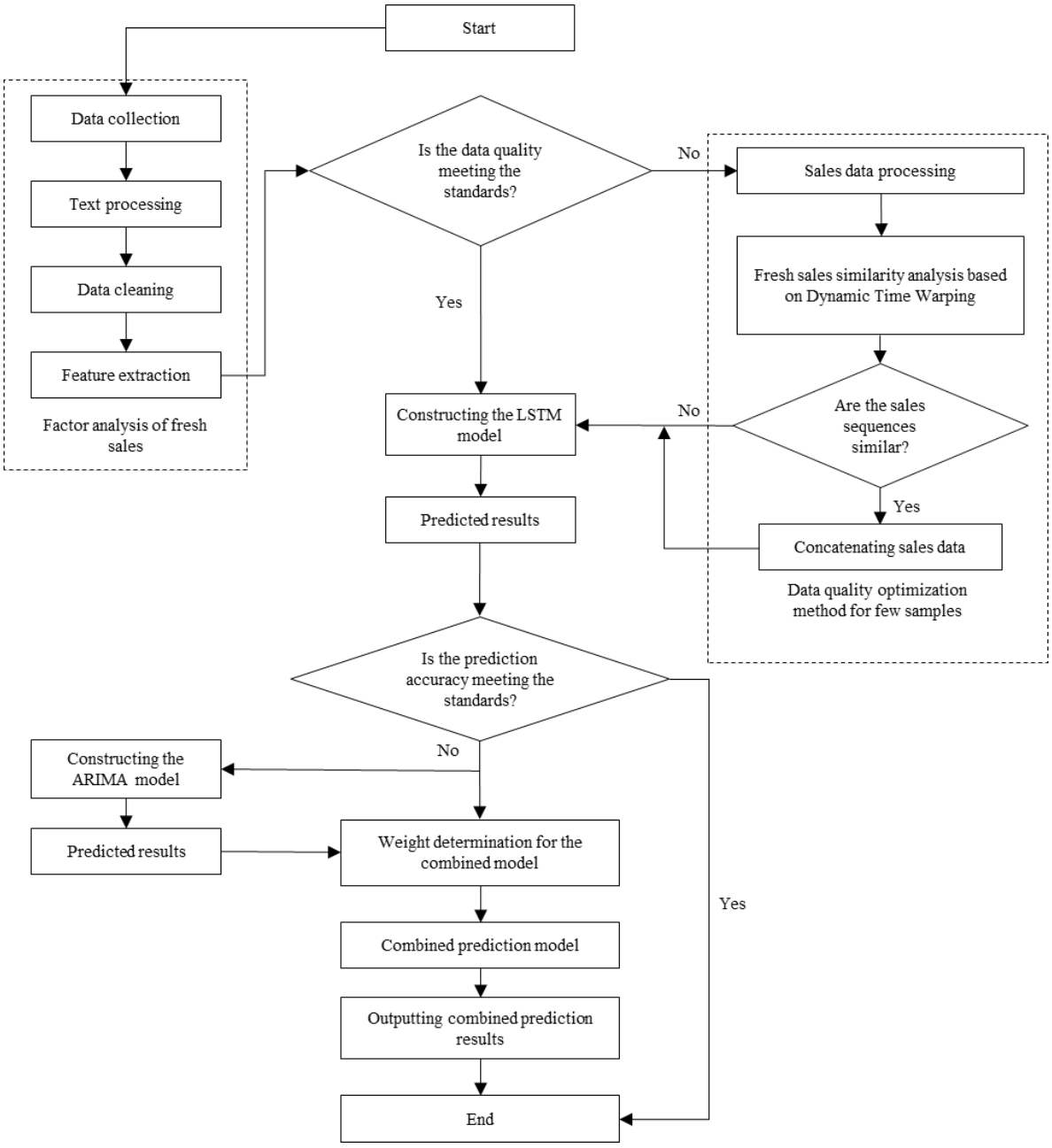

Figure 2. Flowchart of the Fresh Produce Sales Prediction Framework

3.4. Method of Constructing the ARIMA-LSTM Combined Prediction Model

The concept of a combined model involves using multiple single models to make predictions for a target problem and then appropriately weighted averaging of the prediction results to enhance prediction accuracy [27]. Assuming that for a given prediction problem, the true observations at a certain time are denoted as $\{y_t\}$, where $t \in [1,n]$, and a is the number of single prediction model used in the combined model, with $a \geq 2$. Let y_{it}, where $i \in [1,a]$, represent the prediction values of the i-th single prediction model at time t. Let w_i, where $i \in [1,a]$, denote the weights assigned to the prediction results of the i-th single model, and these weights satisfy:

$$\sum_{i=1}^{a} w_i = 1, w_i \geq 0. \quad (9)$$

The prediction value of the combined model can then be expressed as:

$$\hat{y}_t = \sum_{i=1}^{a} w_i y_{it}. \quad (10)$$

By analyzing the time series plots of daily sales data for fresh produces in supermarkets in Hubei province, we observed that the sales data exhibit strong fluctuations within local sequences and display cyclic patterns within a

certain range, indicating nonlinear characteristics. Additionally, there is an overall trend, suggesting linear characteristics. While in most cases, the LSTM model has achieved good prediction accuracy, for certain produces, whether using the LSTM or ARIMA model, the prediction accuracy remains mediocre.

The ARIMA model builds time series models based solely on the historical data itself, without considering external influencing factors as model inputs. Therefore, while it fits linear data well, it cannot identify nonlinear features within the sample sequence. Consequently, if external factors change the development trend within the sample sequence, traditional time series models cannot provide accurate predictions. However, daily sales of fresh produces are influenced not only by their own sales and price factors but also to a large extent by external factors. There are numerous influencing factors with complex interactions, often leading to unstable fluctuations in daily sales time series plots. In comparison to traditional time series models, LSTM models, as a classic type of neural network model, possess strong capabilities for extracting nonlinear features and are suitable for fitting relatively complex nonlinear sample sequences.

Given the complexity of daily sales data for fresh produces and considering the respective advantages of the ARIMA model and LSTM model, this paper proposes the ARIMA-LSTM Combined Prediction Model based on both the ARIMA model and LSTM model, referred to as the ARIMA-LSTM Combined Prediction Model. In cases where the individual predictive performance of the two single models is inferior, this ARIMA-LSTM Combined Prediction Model can enhance the predictive accuracy of the single models.

The output results of the ARIMA-LSTM Combined Prediction Model are obtained by weighted averaging of the predictive results from various single models. There are various methods for determining the weights of the single models within the combined model, and the choice of weight determination method depends on the predictive performance of different models.

In this paper, the weight determination method employed is the Weighted Reciprocal of Errors Averaging Method. This method quantitatively calculates the weights of the single models within the combined model based on the numerical values of error metrics derived from the predictions of the single models. These error metrics objectively reflect the predictive performance of each model.

Assuming there are a total of n different single models within the ARIMA-LSTM Combined Prediction Model, and μ_t represents the numerical value of the error metric for the *i*-th individual model, the weight calculation formula for the *i*-th model is as follows:

$$w_i = \frac{\mu_i^{-1}}{\sum_{i=1}^{n} \mu_i^{-1}}. \quad (11)$$

The weighted reciprocal of errors averaging method considers both the relative size of prediction errors from each single model and quantifies the specific variations in prediction errors during the weight determination process. This means that when computing the final combined model, models with smaller errors will be assigned larger weights, while models with larger errors will have smaller weights, in order to more accurately reflect each model's contribution to the final result. This approach is particularly suitable as a weight determination rule for the ARIMA-LSTM Combined Prediction Model proposed in this paper.

4. Experiment

4.1. Datasets

The data used in this study were sourced from the sales records of major supermarkets in a city in Hubei Province, China. The time span covered the period from January 1, 2017, to August 1, 2021. These supermarkets were categorized into four classes based on their size, denoted as A, B, C, and D. Class A represents large shopping malls and has the most comprehensive data. Therefore, this paper primarily focused on Class A stores, which comprised eight stores. To manage the length of this paper, data from 4 of these Class A stores were predominantly used. These Class A stores collectively sold 3,050 types of fresh produce, resulting in a total of 2,227,359 sales records. All the data mentioned in this paper were stored and retrieved using SQL Server database.

The primary data source for this paper is sales data, where each record represents a single transaction for a specific produce. Therefore, each record corresponds to a single type of produce. It's important to note that the sales amount does not represent the unit price of the produce; rather, it varies with the quantity sold. As a result, the unit price of each product is calculated by dividing the sales amount by the quantity sold. The initial database structure provided by the stores is illustrated in Table 1.

Table 1. Sales Data Table

Field Name	Field Type	Field Description
id	int	Database auto-increment id
store_code	varchar(24)	A numerical code or identifier for the store.
sales_date	varchar(20)	The date of the sale in the format "yyyy-mm-dd."
goods_code	varchar(15)	A unique identifier for the produce
sales_amount	float	The amount of the produce sold.
sales_income	float	The total income of the sale

The store data table records information about the store's type and address, with its specific database structure in Table 2.

Table 2 Store Data Table

Field Name	Field Type	Field Description
id	int	Database auto-increment id
store_code	varchar(24)	A numerical code or identifier for the store.
store_name	varchar(100)	Store's name
store_address	varchar(200)	Store's address
store_type	varchar(10)	Store's type

Not only sales data but also factors such as day of the week, holiday information, and weather data can have an impact on sales prediction. Therefore, this paper needs to collect and process these data, and the following operations are performed in a Python environment.

For day of the week information, this paper converts the sales date information to a DateTime object and directly use the weekday () method available in Python's DateTime object to obtain the day of the week as an integer.

For holiday information, we use a third-party library called "chinese_calendar" and call the appropriate method to obtain holiday information. This library includes Chinese statutory holidays and some significant festivals, such as Valentine's Day, Lantern Festival, and Children's Day. The obtained information will be in the form of a string containing the holiday name.

Historical weather information needs to be obtained through web scraping. We scrape data from the "2345 Weather Forecast" website. The basic principle of web scraping is to use a combination of city IDs and year-month values to access historical weather details pages for a specific city. By running a web scraping program with city IDs and all year-month combinations, we can automatically retrieve historical weather information for that city. The collected information may include daily minimum and maximum temperatures, weather conditions, wind direction, wind speed, and pollution levels. We link this historical weather data to store locations using the store address information from Table 2. The database structure for historical weather data is shown in Table 3.

Table 3 Historical Weather Data Table

Field Name	Field Type	Field Description
id	int	Database auto-increment id
weather_date	varchar(20)	Date，formatted as yyyy-mm-dd
l_temp	varchar(20)	Min Temperature, formatted as x℃
h_temp	varchar(20)	Max Temperature, formatted as x℃
weather	varchar(50)	Weather
wind_dir	varchar(50)	Wind direction
wind_power	int	Wind speed
aqi	int	Pollution level
city	varchar(50)	City

4.2. Experimental setup

In the process of quantitatively evaluating the prediction results and during deep learning gradient descent, it is necessary to select appropriate prediction evaluation metrics. For the numerical prediction problems addressed in this research, assuming the predicted values are $\hat{y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n\}$ and the true values are $y = \{y_1, y_2, \cdots, y_n\}$, the evaluation metrics employed in this paper include the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE):

(1) Mean Absolute Percentage Error, often abbreviated as MAPE, is a metric that measures the percentage difference between errors and true values, based on the Mean Absolute Error. Its formula is as follows:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{12}$$

Here, n represents the number of samples, $\hat{y}_i$ is the predicted value for the i-th sample, and y_i is the corresponding true value.

(2) Root Mean Square Error, often abbreviated as RMSE, is a metric that builds upon the Mean Squared Error by taking the square root of the result. This is advantageous because it provides a more intuitive sense of the magnitude of prediction errors. The formula for RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{13}$$

Here, n represents the number of samples, $\hat{y}_i$ is the predicted value for the i-th sample, and y_i is the corresponding true value.

The 12-dimensional time-series data that influence sales, as mentioned earlier, are presented in Table 4.

Below, one day is taken as the reference, and the 12 influencing factors are measured in days as follows: Last day's sales, Last week's sales, Last year's sales (Solar calendar), Last year's sales(Lunar calendar), Current day's price, Last day's price, Last week's price, Last year's price(Solar calendar), Last year's price(Lunar calendar), Weather information, Holiday information, Day of the week.

This paper needs to transform the relevant data of sales influencing factors into a format friendly to correlation analysis algorithms, such as converting textual information into numerical information and analyzing the rules of conversion.

In text information mining, it is common to apply certain rules to sort textual information to establish connections and patterns among them. In the context of fresh produce sales, weather information has the most significant impact, as adverse weather conditions tend to reduce people's willingness to travel, subsequently affecting the sales of fresh produces.

Table 4 Example Table of Data for 12 Influencing Factors

Factors	Date		
	Jan 1, certain year	Jan 2, certain year	Jan 3, certain year
Last day's sales	269	191	264
Last week's sales	176	122	159
Last year's sales (Solar calendar)	513	454	287
Last year's sales (Lunar calendar)	308	239	297
Current day's price	3.5	3.8	3.8
Last day's price	3.5	3.5	3.8
Last week's price	3.5	3.5	3.5
Last year's price (Solar calendar)	2.5	2.5	2.5
Last year's price (Lunar calendar)	3.5	3.8	3.8
Weather information	0	1	3
Holiday information	1	0	0
Day of the week	0	1	2

In this paper, similar weather conditions are merged based on severity levels. The merged weather information is then sorted and mapped to numerical ranges in a specific order. For example, Sunny, Cloudy, Partly Cloudy, and Foggy are grouped together as they have little impact on people's purchasing behavior regarding fresh produces. On the other hand, Light Rain, Light Snow, and Sleet have a mild impact, with Sleet being a condition that falls between rain and snow and typically involves a combination of light rain and snow. The results are presented in Table 5.

Table 5 Weather Information Mapping Scheme

Weather	Severity Level	Number
Sunny Overcast Partly Cloudy Foggy	None	0
Light Rain Light Snow Sleet	Low	1
Moderate Rain Moderate Snow	Medium	2
Heavy Rain Heavy Snow	High	3
Rainstorm Blizzard	Very high	4

Different from weather information, there is no inherent order among holiday and day of week information. For instance, it's challenging to establish a logical order between holidays like the Spring Festival and National Day. Similarly, there's no progressive relationship between weekdays, such as Monday and Sunday. One suitable approach is to treat each type of information independently, as there's no inherent order among them. In the project code, we use the LabelEncoder object from the sklearn library to map text information to numerical values. For holiday information, we assign unique numerical labels to each holiday, such as marking non-holidays as 0, New Year's Day as 1, Spring Festival as 2, and so on, using distinct numerical values for all holidays. For day of week information, we simply assign numerical labels from 0 to 6 for Monday through Sunday.

4.3. Experimental results of the ARIMA-LSTM Combined Prediction Model

4.3.1. Combined model with superior single-model prediction accuracy

In this experiment, when both the ARIMA model and LSTM model exhibit relatively superior prediction accuracy, this paper utilizes the Weighted Reciprocal of Errors Averaging Method as the weight determination approach for the ARIMA-LSTM Combined Prediction Model. For the daily cucumber sales at a specific A-class store, the single model's MAPE and weight percentages are presented in Table 6.

Table 6 Weight Table for Superior Single Model Predictions

	ARIMA model	LSTM model
MAPE	31.63%	15.14%
Weight	0.32	0.68

The prediction results of the combined model are shown in Figure 3.

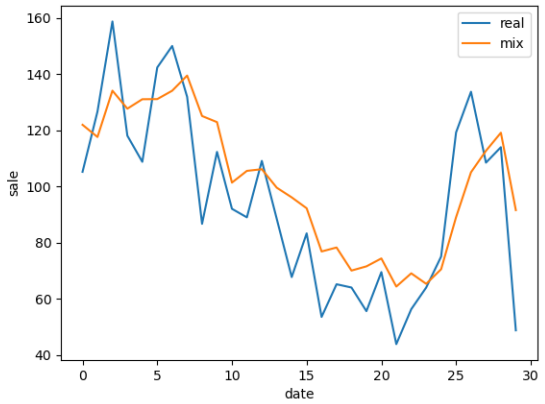

Figure 3. Combined Model Prediction Chart with Superior Single Model Predictions

The MAPE value for this ARIMA-LSTM Combined Prediction Model is 19.05%. It can be observed that the combined model, even with superior individual models, does not show a significant improvement in prediction accuracy. In fact, it exhibits a slightly lower accuracy compared to the LSTM model alone. To understand the reasons behind this, this paper conducts a comparison of predictions between the two single models and the combined model, as shown in Figure 4.

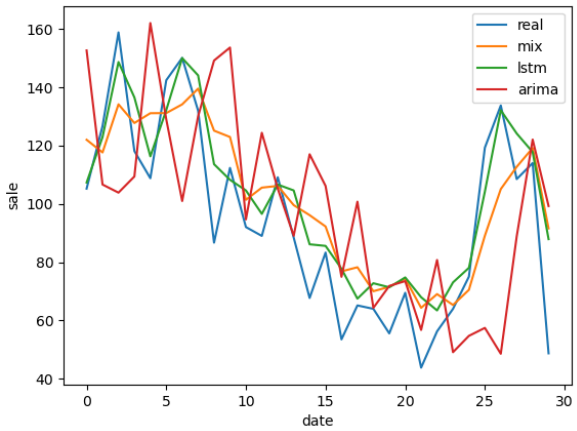

Figure 4. Comparison Chart of Superior Single Model Predictions

From Figure 4, it can be observed that the ARIMA model exhibits noticeable lag. While it provides more accurate predictions at certain time points (such as on the 21st day) compared to the LSTM model, in most cases, the LSTM model performs better in terms of accuracy. The ARIMA model tends to accumulate prediction errors in the combined model, resulting in the ARIMA-LSTM Combined Prediction Model's accuracy being lower than that of the LSTM model for most of the time. This indicates that the prediction of daily sales of fresh produce is a complex problem. While extensive research and empirical evidence suggest that combined prediction can improve the accuracy of single models, the specific problem still requires individual analysis. It's not the case that combined models outperform single models in all specific scenarios.

4.3.2. Combined model with inferior single-model prediction accuracy

The previous experiment illustrated that when the single model's prediction accuracy is relatively superior, the ARIMA-LSTM Combined Prediction Model can lead to negative optimization. In this experiment, we chose situations where the single model's prediction accuracy, especially the LSTM model, was relatively inferior. We observed the changes in prediction accuracy of the ARIMA-LSTM Combined Prediction Model. We established the ARIMA-LSTM Combined Prediction Model for the daily sales of premium bananas in a certain Class A store, and the MAPE and weight percentages of the single models are shown in Table 7.

Table 7 Weight Table for Inferior Single Model Predictions

	ARIMA model	LSTM model
MAPE	45.43%	30.81%
Weight	0.40	0.60

The prediction results of the combined model are shown in Figure 5:

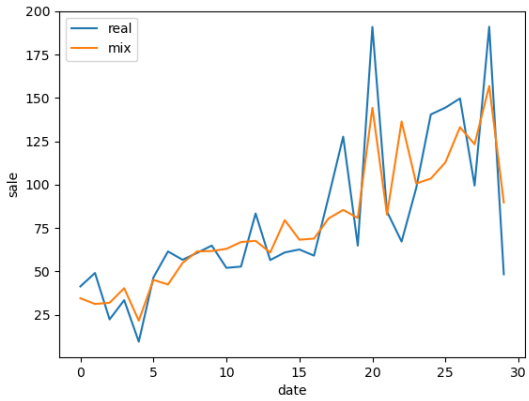

Figure 5. Combined Model Prediction Chart with Inferior Single Model Predictions

The model has a MAPE value of 26.95%, showing an improvement compared to the ARIMA model with a MAPE value of 45.43% and the LSTM model with a MAPE value of 30.81%. Relative to LSTM, the improvement is 3.86%. From Figure 5, it can be observed that the ARIMA-LSTM Combined Prediction Model, except for a few days (such as the 22nd day), generally captures the accurate daily sales trends and variations. Figure 6 provides a comparison of the predictions of each model.

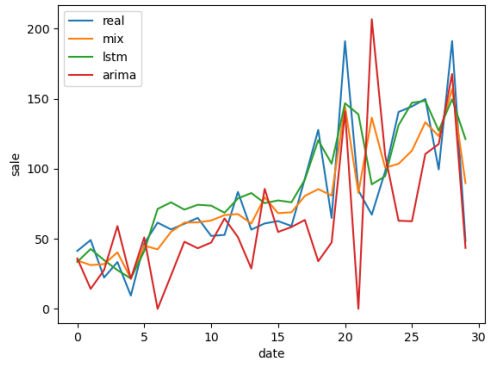

Figure 6. Comparison Chart of Inferior Single Model Predictions

From Figure 6, it can be observed that although the ARIMA model still exhibits some lag in its predictions at certain times, it can promptly reflect the changes in daily sales trends at other times, such as between the 20th and 23rd days. During this period, the LSTM model shows a lag in its predictions. However, for the rest of the time, the LSTM model fits the data better than the ARIMA model. Therefore, the ARIMA model and the LSTM model can complement each other, resulting in more accurate predictions in the ARIMA-LSTM Combined Prediction Model.

Based on these two experiments, some initial conclusions can be drawn: In the prediction of daily sales of fresh produces, if the LSTM model achieves high prediction accuracy, it can be directly used as the final model for predicting daily sales. If the LSTM model's prediction accuracy is low, the ARIMA-LSTM Combined Prediction Model can be used as the final model for predicting daily sales of fresh produces.

4.4. Comparative experiments of the Fresh Produce Sales Prediction Framework

To validate the superiority of the Fresh Produce Sales Prediction Framework, this section selected the time series prediction framework Prophet, as well as ARIMA and traditional LSTM models, as control groups. The data from four A-class stores were chosen for the experiment.

To compare the prediction performance between the Prophet framework and the proposed Fresh Produce Sales Prediction Framework, the study trained the model on cucumber fresh produce data from the four A-class stores with the most comprehensive data. The target was the daily sales for the next 30 days. The results were then compared with the predictions generated by LSTM models and ARIMA models that used only sales data as features. The evaluation metrics used for comparison were MAPE and RMSE.

Table 8 shows the comparative results of the prediction performance of the Prophet framework, ARIMA model, LSTM model, and the proposed model for cucumber sales at A, B, C, and D-class stores.

Table 8 Comparative Analysis of the Performance of Four Time Series Prediction Models

	Prophet		ARIMA		Traditional LSTM		FPSPF	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Store A	360.91%	109.93	606.79%	123.69	411.42%	98.84	**21.25%**	**52.62**
Store B	50.81%	40.71	48.44%	55.27	35.91%	43.58	**10.28%**	**14.09**
Store C	42.90%	38.15	31.63%	35.82	27.41%	29.54	**11.54%**	**14.44**
Store D	47.43%	50.77	44.74%	80.59	25.38%	60.60	**15.95%**	**43.24**

5. Conclusion

This paper begins with a specific problem of predicting daily sales of fresh produces and conducts an analysis of the current research status in this domain. We propose the Fresh Produce Sales Prediction Framework that combines multidimensional time-series data affecting store sales. This model accurately predicts future daily sales of fresh produces over a certain period, providing replenishment recommendations for stores. Consequently, it helps reduce fresh produce waste caused by subjective factors and enhances business profits. Comparative experimental results indicate the outstanding predictive performance of the model proposed in this paper.

Furthermore, this paper employs real-life daily sales data of fresh produces to construct and predict using three different models: the ARIMA time series model, the LSTM neural network model, and the ARIMA-LSTM Combined Prediction Model. The prediction performance of the combined model is compared with that of the single models in scenarios where single model predictions range from inferior to superior. The results show that in cases where single model predictions are inferior, the combined model performs better. This validates that the ARIMA-LSTM Combined Prediction Model proposed in this paper outperforms single models in certain situations.

Acknowledgements

The project was supported by Hubei Key Laboratory of Big Data in Science and Technology (Wuhan Library of Chinese Academy of Science) (No. ZK2022003).

References

[1] C.-H. W .Considering economic indicators and dynamic channel interactions to conduct sales forecasting for retail sectors[J].Computers & Industrial Engineering, 2022:165.

[2] Zhou X, Zhai N, Li S, et al. Time series prediction method of industrial process with limited data based on transfer learning[J]. IEEE Transactions on Industrial Informatics, 2023,19 (5):6872-6882

[3] Ren L, Jia Z, Laili Y, et al. Deep Learning for Time-Series Prediction in IIoT: Progress, Challenges, and Prospects[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.

[4] Bai Y T, Jia W, Jin X B, et al. Nonstationary Time Series Prediction Based on Deep Echo State Network Tuned by Bayesian Optimization[J]. Mathematics, 2023, 11(6): 1503.

[5] Novianti Trisita and Utami Issa Dyah and Prima Dania Wike Agustin. Forecasting of salt demand using ARIMA model to prevent the bullwhip effect in salt supply chain[J]. Journal of Physics: Conference Series, 2022, 2193(1).

[6] Longzhou Chen. Risk Analysis of Pork Market Price in China based on ARIMA Model[J]. Scientific Journal of Economics and Management Research, 2022, 4(2).

[7] Zhao Xin. Forecast Port Container Throughput based on ARIMA Model[J]. Frontiers in Economics and Management, 2022, 3(2).

[8] Arunraj N S, Ahrens D. A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting[J]. International Journal of Production Economics, 2015, 170: 321-335.

[9] Tofik Mussa Reshid. Ethiopian Economic Features and Changing Persistence: A Time Series Analysis. 2020, 9(3).

[10] David Adugh Kuhe,Thomas Akwana Obed.An ARMA Model for Short-term Prediction of Hepatitis B Virus Seropositivity among Blood Donors in Lafia-nigeria. 2019 , : 1-11.

[11] Prathap Rudra Boppuru, Ramesha K..Spatio-temporal Crime Analysis Using KDE and ARIMA Models in the Indian Context[J]. International Journal of Digital Crime and Forensics (IJDCF), 2020,12(4).

[12] Ahmad T E et al. Fisheries forecasting, physical approach comparison between regression and autoregressive integrated moving average (ARIMA)[J]. IOP Conference Series: Earth and Environmental Science, 2022, 967(1).

[13] Posch K, Truden C, Hungerländer P, et al. A Bayesian approach for predicting food and beverage sales in staff canteens and restaurants[J]. International Journal of Forecasting, 2022, 38(1): 321-338.

[14] Hossain M S , Mahmood H . Short-term photovoltaic power forecasting using an LSTM neural network andsynthetic weather forecast[J].IEEE Access,2020,8:172524-172533.

[15] Siami-namini S,Tavakoli N , Namin A S.A Comparison of ARIMA and LSTM in Forecasting Time Series[C]2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) IEEE , 2018:1394 -1901.

[16] Lee W I, Chen C W, Chen K H, et al. A comparative study on the forecast of fresh food sales using logistic regression, moving average and BPNN methods[J]. Journal of Marine Science and Technology, 2012, 20(2): 4.

[17] Absar Nurul et al. The efficacy of deep learning based LSTM model in forecasting the outbreak of contagious diseases[J]. Infectious Disease Modelling, 2021.

[18] Shahani Niaz Muhammad et al. Predictive modeling of drilling rate index using machine learning approaches: LSTM, simple RNN, and RFA[J]. Petroleum Science and Technology, 2022, 40(5) : 534-555.

[19] Pei Ying et al. Trend Prediction of DC Measuring System Based on LSTM[J]. Journal of Physics: Conference Series, 2021, 2083(3).

[20] Wang L, Li X, Zhu H, et al. Influencing factors of livestream selling of fresh food based on a push-pull model: A two-stage approach combining structural equation modeling (SEM) and artificial neural network (ANN)[J]. Expert Systems with Applications, 2023, 212: 118799.

[21] Bates J M, Granger C W J. The combination of forecasts[J].Journal of the Operational Research Society,1969,20(4):451-468.

[22] Chandriah K K, Naraganahalli R V. RNN/LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting[J]. Multimedia Tools and Applications, 2021, 80(17): 26145-26159.

[23] Guo Yan et al. Research on Precipitation Forecast Based on LSTM–CP Combined Model[J]. Sustainability, 2021, 13(21) : 11596-11596.

[24] Luo T, Chang D, Xu Z. Research on Apparel Retail Sales Forecasting Based on xDeepFM-LSTM Combined Forecasting Model[J]. Information, 2022, 13(10): 497.

[25] Wang S, Yang Y. M-GAN-XGBOOST model for sales prediction and precision marketing strategy making of each product in online stores[J]. Data Technologies and Applications, 2021, 55(5): 749-770.

[26] Ahmed S, Chakrabortty R K, Essam D L, et al. Poly-linear regression with augmented long short term memory neural network: Predicting time series data[J]. Information Sciences, 2022, 606: 573-600.

[27] Bekiroglu K, Gulay E, Duru O. A multi-method forecasting algorithm: Linear unbiased estimation of combine forecast[J]. Knowledge-Based Systems, 2022, 239: 107990.