Research on an Integrated Model for Logistics Demand Forecasting Based on Random Forests and LSTM

Yiming Xu †,*, Ruiqi Zhou †, Wanyu Tang †

Delaware Data Science Institute, Southwestern University of Finance and Economics, Chengdu, China

* Corresponding Author Email: 3315562147@qq.com

†.These authors also contributed equally to this work

Abstract. This paper proposes a logistics demand forecasting model integrating Support Vector Machines (SVM), Random Forests, and Long Short-Term Memory (LSTM) networks, focusing on the construction logic of this ensemble model and methods for integrating prediction results. First, SVM maps raw features into a high-dimensional space, combining sparse linear criteria for feature selection to preserve nonlinear representation capabilities while controlling model complexity. Second, enhanced features and lagged terms are fed into both the random forest and LSTM. The random forest enhances robustness through multi-tree voting/averaging, while the LSTM captures temporal dependencies via its gated structure, ensuring information consistency. Finally, prediction results are adaptively weighted and integrated using an inverse error criterion based on model performance on the validation set, followed by parameter tuning and cross-validation. This model supports logistics inventory management decisions by integrating the strengths of three algorithms. It effectively handles nonlinear and time-series data, enhances prediction accuracy and generalization capabilities, and demonstrates strong robustness against outliers and missing values.

Keywords: Support vector machine; random forest; long short-term memory; ensemble learning model.

1. Introduction

This paper focuses on practical issues in the logistics industry, such as inefficient inventory management and resource waste caused by insufficient demand forecasting accuracy. It aims to construct a comprehensive forecasting model by integrating Support Vector Machines (SVM), Random Forests, and Long Short-Term Memory (LSTM) networks, thereby providing scientific basis for inventory decision-making in logistics enterprises [1][2]. First, addressing the complexity of logistics demand influenced by regional, seasonal, weather, promotional, and other factors, SVM maps raw demand features into a high-dimensional space. Combined with sparse linear criteria, this approach filters key features, preserving nonlinear data correlations while reducing computational redundancy in subsequent models [3][4]. Second, the processed enhanced features and demand time-lagged components are simultaneously fed into two complementary submodels: Random Forest enhances robustness against anomalous demand data through multi-tree voting and averaging mechanisms; LSTM utilizes gated structures to accumulate historical demand states, precisely capturing dynamic patterns of demand evolution over time [5]. Finally, based on the prediction errors of each submodel on the validation set, results are integrated through adaptive weighting using the inverse error criterion. Parameter tuning and cross-validation ensure the model's stability and applicability across diverse logistics scenarios [6].

Experimental results demonstrate that the constructed integrated model significantly outperforms standalone SVM, Random Forest, or LSTM models in logistics demand forecasting accuracy. It exhibits strong generalization capabilities across different regions and product categories, effectively accommodating demand fluctuations caused by variables such as promotions and weather. This model provides a reliable basis for dynamic adjustments to logistics inventory and optimized resource allocation.

2. Description of the Integrated Algorithm Principles

To construct a logistics demand forecasting model and optimize inventory management strategies based on forecasting results, algorithms such as Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM) are integrated. This integration fully leverages their unique advantages in processing different types of data and pattern recognition, thereby improving the accuracy of forecasting.

Support Vector Machine (SVM): A machine learning algorithm based on supervised learning, which has the ability to handle nonlinear problems and high-dimensional data.

Random Forest: An ensemble learning method that enhances model stability and accuracy by constructing multiple decision trees, and exhibits good robustness to outliers and missing values.

Long Short-Term Memory (LSTM): A special type of recurrent neural network, specifically designed to handle and predict long-term dependencies in time series data, and performs excellently in time series forecasting.

The detailed principle description of this ensemble learning method is as follows:

2.1. Feature Extraction and Selection Based on SVM

To characterize the nonlinear relationships in the demand sequence, the original feature $\mathbf{x}_t \in \square^p$ is first mapped to a high-dimensional feature space, and then concatenated with the original features to form an augmented representation for downstream models:

$$\mathbf{z}_t = \left[\mathbf{x}_t, \phi(\mathbf{x}_t)\right], \quad \phi(\mathbf{x}) = \left(\kappa(\mathbf{x}, \mathbf{v}_1), \ldots, \kappa(\mathbf{x}, \mathbf{v}_m)\right), \tag{1}$$

where $\kappa(\cdot,\cdot)$ is a kernel function (such as RBF), and $\{\mathbf{v}_j\}_{j=1}^m$ are representative anchor points (obtained from training samples or clustering centers). To avoid redundancy and enhance interpretability, a sparse linear criterion is employed on the validation set to perform feature selection on $\mathbf{z}_t$:

$$\min_\beta \frac{1}{|\mathrm{V}|}\sum_{t\in \mathrm{V}}\left(y_t - \beta^{\bullet}\,\mathbf{z}_t\right)^2 + \lambda\|\,\beta\|_1, \tag{2}$$

The obtained non-zero coefficients correspond to "retained features". In this way, the nonlinear characterization capability of SVM is utilized, and the complexity is controlled through sparsification, reducing interference to subsequent models. These features can capture complex patterns and relationships in the data, providing a foundation for further processing by subsequent models.

2.2. Model Integration and Fusion Based on Random Forest and LSTM

After obtaining the augmented feature $\mathbf{z}_t$, it is input into two types of complementary models together with the necessary lag terms:

(1) Random Forest (RF) improves robustness through voting/averaging of multiple sub-models:

$$\hat{y}_{t+h}^{\mathrm{RF}} = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{z}_t), \tag{3}$$

where $T_b(\cdot)$ is the regression output of the b-th tree.

(2) LSTM models temporal dependencies, utilizing a gating structure to accumulate states and output predictions:

$$\mathbf{c}_t = \mathbf{f}_t \square\ \mathbf{c}_{t-1} + \mathbf{i}_t \square\ \tilde{\mathbf{c}}_t, \quad \mathbf{h}_t = \mathbf{o}_t \square\ \tanh(\mathbf{c}_t), \quad \hat{y}_{t+h}^{\mathrm{LSTM}} = g(\mathbf{h}_t). \tag{4}$$

The two respectively target "nonlinear mapping" and "temporal dependence", being naturally complementary; using $\mathbf{z}_t$ as a unified input can ensure information consistency, facilitating comparison and fusion.

By taking the features extracted by SVM together with other original features as inputs to Random Forest and LSTM:

Random Forest enhances the model's stability and generalization ability by constructing multiple decision trees and integrating their prediction results.

Meanwhile, LSTM can capture historical information in time series data, understand the changing trends of data over time, and predict future demand changes.

2.3. Adaptive Weight Integration of Forecasting Results

This paper adaptively assigns weights according to the performance of each model on the validation set V to obtain the final prediction:

$$\hat{y}_{t+h} = w_{\mathrm{RF}}\hat{y}_{t+h}^{\mathrm{RF}} + w_{\mathrm{LSTM}}\hat{y}_{t+h}^{\mathrm{LSTM}}, \quad w_{\mathrm{RF}} + w_{\mathrm{LSTM}} = 1, w_{\mathrm{RF}}, w_{\mathrm{LSTM}} \geq 0. \tag{5}$$

The weights are given by the following simple and robust inverse error criterion (to avoid overfitting and facilitate implementation):

$$w_k = \frac{e_k^{-1}}{\sum_j e_j^{-1}}, \quad e_k = \frac{1}{|\mathrm{V}|}\sum_{t\in\mathrm{V}}\left|y_t - \hat{y}_t^{(k)}\right|, \quad k \in \{\mathrm{RF}, \mathrm{LSTM}\}. \tag{6}$$

This integration method can fully leverage the advantages of each model and improve the accuracy of forecasting.

2.4. Optimization and Evaluation of the SVM-RF-LSTM Integrated Model

Parameter tuning and cross-validation are performed on the ensemble model to ensure its stability and generalization ability across different datasets. Meanwhile, comparative experiments and evaluation metrics (such as accuracy, recall, F1-score, etc.) are used to verify the superiority of the ensemble model over single models.

In summary, through the aforementioned ensemble learning method, this study aims to construct a forecasting model that can comprehensively capture complex patterns in logistics demand data and enhance forecasting accuracy and generalization ability. This model can fully utilize the nonlinear processing capability of SVM, the stability and generalization ability of Random Forest, and the time-series data processing capability of LSTM, thereby providing a scientific basis for decision support in the logistics industry.

3. Experimental Results and Analysis

To evaluate the effectiveness of the proposed method, a synthetic logistics demand dataset simulating real business scenarios is constructed. Considering that real-world demand is often influenced by the combined effects of multiple nonlinear factors (e.g., region, product category, season, promotion, weather), this study adopts Random Forest as an implicit data generation mechanism to generate logistics demand based on specific input features. On this basis, controlled random perturbations are added to reflect uncertainty and noise in actual operations.

3.1. Generation Mechanism

Let $\mathbf{x}_t$ include date-related features (e.g., trend, seasonal components), region, product type, weather, promotion, and other variables. The demand is generated by the following formula:

$$y_t = F_{\mathrm{RF}}(\mathbf{x}_t) + \varepsilon_t \tag{7}$$

where F_{RF} denotes a Random Forest regressor obtained through preset rules; ε_t is a random perturbation term, used to simulate measurement errors, short-term shocks, and unobserved factors.

To make the noise amplitude and structure more consistent with the real-world characteristic of "fluctuations varying by scenario", an annealing-based noise injection strategy is adopted. For a given temperature sequence, the noise scale is gradually reduced—ensuring higher explorability in the early stage and greater alignment with stable business states in the later stage. The mathematical expression is as follows:

$$\varepsilon_t^{(s)} \sim \mathrm{N}\left(0, \sigma^2\left(\tau_s\right)\right), \quad \sigma\left(\tau_s\right) = \sigma_0 \cdot \tau_s, \quad \tau_{s+1} < \tau_s. \tag{8}$$

The final demand is determined by $y_t = F_{\mathrm{RF}}\left(\mathbf{x}_t\right) + \varepsilon_t^{(S)}$. This process draws on the idea of "temperature decrease and gradual convergence" from simulated annealing, which is used to control noise intensity and simulate real-world uncertainty—rather than solving an optimization problem.

3.2. Features and Rules

To enhance flexibility and interpretability, $\mathbf{x}_t$ includes the following components, with their influence directions or magnitudes defined by simple business rules:

1. Date and Trend: Linear/piecewise trends and weekly-monthly-annual seasonal components (e.g., intra-week cycles, holiday indicator variables).
2. Region and Product Category: Sampled from limited enumeration sets, allowing the setting of baseline demand levels and elasticity differences for different regions/categories.
3. Weather: Indicators such as temperature/precipitation (generated by a random process in the example; driven by historical data in practical applications).
4. Promotion: Binary or intensity-based variables, used to trigger short-term demand surges.

In specific implementation:

1. An initial "weak-logic" demand is first synthesized based on the above rules;
2. This initial value is used as the label to fit F_{RF}, enabling the Random Forest to "reconstruct" nonlinear relationships in a higher-dimensional combinatorial space;
3. Finally, annealing-based noise is added to obtain the final y_t.

This approach not only retains the interpretability of business rules but also introduces model-driven complex interaction terms.

3.3. Generation Process

1. Time Axis and Primary Keys: Generate date indices and combined keys of (region, product category);
2. Feature Construction: Extract features such as trend, seasonal components, holidays, weather, and promotion to form $\mathbf{x}_t$;
3. Weak-Logic Initial Value: Generate the initial demand $\tilde{y}_t$ based on business intuition (e.g., promotion day increments, off-peak/peak season coefficients);
4. Forest Fitting: Fit F_{RF} using $\left(\mathbf{x}_t, \tilde{y}_t\right)$ to enrich nonlinearity and interactions;
5. Annealing Noise: Set the temperature sequence τ_s, inject $\varepsilon_t^{(s)}$, and obtain the final y_t;
6. Data Storage: Output the synthetic dataset containing all features and y_t.

The final generated dataset is presented in Table 1.

Table 1. Logistics demand dataset

Date	Region	ProductType	Demand	Weather	Promotion
01-01	Region3	TypeB	36	Snowy	0
01-02	Region1	TypeC	127	Snowy	0
01-03	Region3	TypeB	118	Snowy	0
01-04	Region3	TypeC	112	Sunny	0
01-05	Region1	TypeB	113	Sunny	1
...	...	...	...	...	...
12-27	Region3	TypeA	105	Sunny	0
12-28	Region2	TypeA	153	Sunny	0
12-29	Region2	TypeB	134	Snowy	0
12-30	Region3	TypeC	61	Snowy	0
12-31	Region2	TypeA	75	Rainy	0

In the constructed simulated logistics dataset, statistical analysis is conducted on the overall distribution of demand volume to verify its rationality and representativeness. In the descriptive statistics of the demand data:

The first quartile (Q1) is 87; The median (Q2) is 118; The third quartile (Q3) is 145; The maximum value is 238. This indicates that most of the demand is concentrated in the interval of 87–145, and the median is close to the overall mean. Additionally, there are extreme values up to 238, which shows the data exhibits a certain degree of volatility and a long-tail characteristic. The distribution is illustrated in Fig. 1.

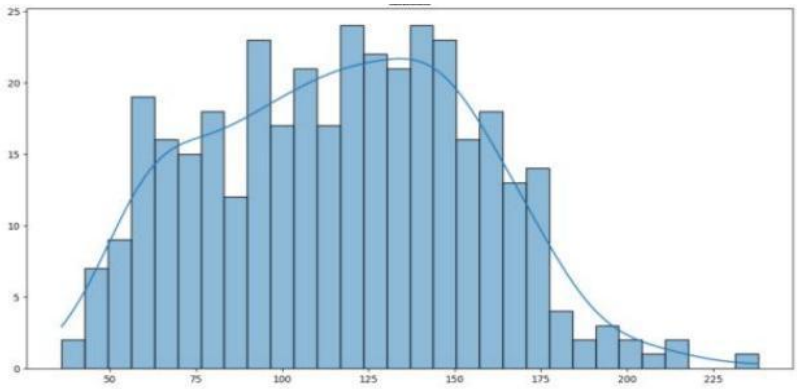

Fig 1. histogram of logistics demand data distribution

The correlation between demand and promotional activities is illustrated in Fig. 2:

Fig 2. Correlation coefficient matrix between demand and promotion

As shown in Fig. 2, there is a positive correlation to a certain extent between Demand and Promotion, with a correlation coefficient of 0.36. This indicates that the demand volume generally

shows an upward trend when promotional activities occur, but the correlation is not absolute. This also suggests that besides the promotion factor, demand is jointly influenced by multiple factors such as season, region, and product type.

The average demand under different regions, product types, and weather conditions obtained finally is presented in Table 2:

Table 2. Average demand across different regions, product types, and weather conditions

Region	ProductType	Weather	Demand
0	Region1	TypeA	Rainy
1	Region1	TypeA	Snowy
2	Region1	TypeA	Sunny
3	Region1	TypeB	Rainy
4	Region1	TypeB	Snowy
5	Region1	TypeB	Sunny
6	Region1	TypeC	Rainy
7	Region1	TypeC	Snowy
8	Region1	TypeC	Sunny
9	Region2	TypeA	Rainy
10	Region2	TypeA	Snowy
11	Region2	TypeA	Sunny
12	Region2	TypeB	Rainy
13	Region2	TypeB	Snowy
14	Region2	TypeB	Sunny
15	Region2	TypeC	Rainy
16	Region2	TypeC	Snowy
17	Region2	TypeC	Sunny
18	Region3	TypeA	Rainy
19	Region3	TypeA	Snowy
20	Region3	TypeA	Sunny
21	Region3	TypeB	Rainy
22	Region3	TypeB	Snowy
23	Region3	TypeB	Sunny
24	Region3	TypeC	Rainy
25	Region3	TypeC	Snowy
26	Region3	TypeC	Sunny

4. Conclusion

This paper proposes an integrated logistics demand forecasting model that combines Support Vector Machines (SVM), Random Forests, and Long Short-Term Memory (LSTM) networks. The core advantage of this model lies in its ability to integrate the characteristics of these three algorithms, enabling it to handle multi-factor driven nonlinear demand relationships while accurately capturing temporal patterns. It also possesses strong robustness, providing direct technical support for inventory management decisions in logistics enterprises. First, the model employs SVM to map raw demand features (such as region, season, weather, and promotional information) into a high-dimensional space. Combined with sparse linear criteria, this process filters key features, preserving core data correlations while avoiding redundant features that increase computational costs, thereby establishing a robust data foundation for the forecasting stage. Second, the processed enhanced features and demand time-lagged components are simultaneously fed into twin models: Random Forest employs a multi-decision tree voting/averaging mechanism to effectively resist interference from abnormal demand data (e.g., sudden surges caused by promotions), while LSTM utilizes a gated structure to accumulate historical demand states, clearly capturing dynamic trends over time. These two models complement each other functionally. Then, based on the prediction errors of each sub-model on the

validation set, weights are adaptively assigned using the inverse error criterion to integrate results, ensuring the final forecast aligns more closely with actual demand. Finally, through parameter tuning and cross-validation, the model's generalization capability across diverse logistics scenarios (e.g., regions, product categories) is enhanced, mitigating risks during enterprise deployment. Future research could further integrate real-time logistics order data to optimize the model's dynamic update efficiency.

References

[1] Sheng Junfan, Zhang Zhiqing. Forecasting Shanghai's Logistics Demand Based on a Combined Forecasting Model [J]. Logistics Science and Technology, 2024, 47(19): 28-32. DOI: 10.13714/j.cnki.1002-3100.2024.19.006.

[2] Zhu Yiding, Zhang Yunchuan, Ma Yunfeng, et al. Multi-dimensional Long Sequence Logistics Demand Forecasting Based on CNN-LSTM-AM Neural Network [J]. Logistics Science and Technology, 2024, 47(18): 49-56+64. DOI: 10.13714/j.cnki.1002-3100.2024.18.010.

[3] Xu Aiping, Hao Yiwei, Zhu Biyun. Research on Medical Supply Inventory Management Based on Hybrid Intelligent Optimization Algorithms [J]. Electronic Design Engineering, 2024, 32(21): 37-40+46. DOI: 10.14022/j. issn1674-6236.2024.21.008.

[4] Yi Jinmei, Zhao Xu, Wei Jingfan. Research on Lanthanum Oxide Price Forecasting Based on a Combined LSTM-SVM Model [J]. Journal of Mudanjiang Normal University (Natural Science Edition), 2025, (02): 11-16+26. DOI: 10.13815/j.cnki.jmtc(ns).2025.02.003.

[5] Huang Chuwen, Guan Yongle, Wang Hongfa. Simulation of Urban Flooding and Water Accumulation Based on the RF-LSTM Model [J]. People's Yellow River, 2025, 47(06): 50-56.

[6] Xia Weihai, Liu Jiali, Feng Fenling. Demand Forecasting for Railway Refrigerated Transportation Based on Random Forest [J]. Journal of Railway Science and Engineering, 2022, 19(04): 909-916. DOI: 10.19713/j.cnki.43-1423/u. T20210517.