

# Research on Logistics Cargo Prediction Based on Machine Learning Methods

Yisheng Gao \*

School of Management, Shandong University of Technology, Zibo, China, 255012

\* Corresponding Author Email: 13210839959@163.com

**Abstract.** Accurate prediction of cargo volume and reasonable arrangement of staff scheduling are crucial for improving the efficiency and punctuality of logistics transport. To address the problem of cargo volume prediction, this paper first conducts a visual analysis of the cargo volume data of sorting centres, and finds that the cargo volume of most sorting centres exhibits a rising trend of fluctuation over time. Then machine learning methods such as Random Forest and SVR are used to construct a prediction model by combining lagging features, trending features and periodic features. After model training and evaluation, the Random Forest model was finally selected to fit the prediction of the cargo volume of different sorting centres, and the rolling prediction strategy was used to ensure the timeliness and accuracy of the prediction results Improved accuracy of logistics cargo forecasting.

**Keywords:** Random Forest, SVR, Cargo Volume Forecasts.

## 1. Introduction

In today's highly competitive logistics market environment, the sorting centre plays a crucial role, undertaking heavy cargo handling and distribution duties. In this context, accurate forecasting of cargo volume and reasonable arrangement of staff become the core elements to ensure the efficiency and punctuality of logistics transport.

For the 56 sorting centres in a logistics network, need to solve the daily and hourly volume forecasts for the sorting centres for the next 30 days, as well as the staff scheduling problems in the sorting centres, as follows:

In terms of logistics volume forecasting, Hu Jiaying [1] conducted qualitative forecasting, linear regression forecasting, time series forecasting and neural network forecasting for express cargo volume, and analysed the advantages and disadvantages of each method. Panli Zhang [2] used MTN method and ARIMA model for order quantity prediction, and combined with immunogenetic algorithm to construct warehouse and storage volume optimisation model. Liu Antelope [3] analysed the status quo of the development of LTL industry and intelligent logistics technology, sorted out the key issues of trunk line optimization, summarized the shortcomings of the current research, and carried out the design of network optimization and outbound scheduling algorithms on this basis. Liu Di [4] used GWO-Elman model for prediction. Firstly, the cargo throughput data of Qingdao Port from January 2010 to December 2019 were collected and preprocessed. Then, the first 75% of the data were used as the training set and the last 25% as the test set. During the training process, the weights and neuron thresholds of the Elman neural network were optimized using the GWO algorithm. Finally, the trained model is applied to the test set to evaluate the prediction performance of the model. Yang Yugee [5] et al. first trained the port cargo throughput data using the Neural Prophet model, obtained the predicted values and calculated the residual series. Then, an LSTM neural network model was established for the residual data to make forecast corrections, and finally reconstructed to obtain the predicted values based on the combined Neural Prophet-LSTM model.

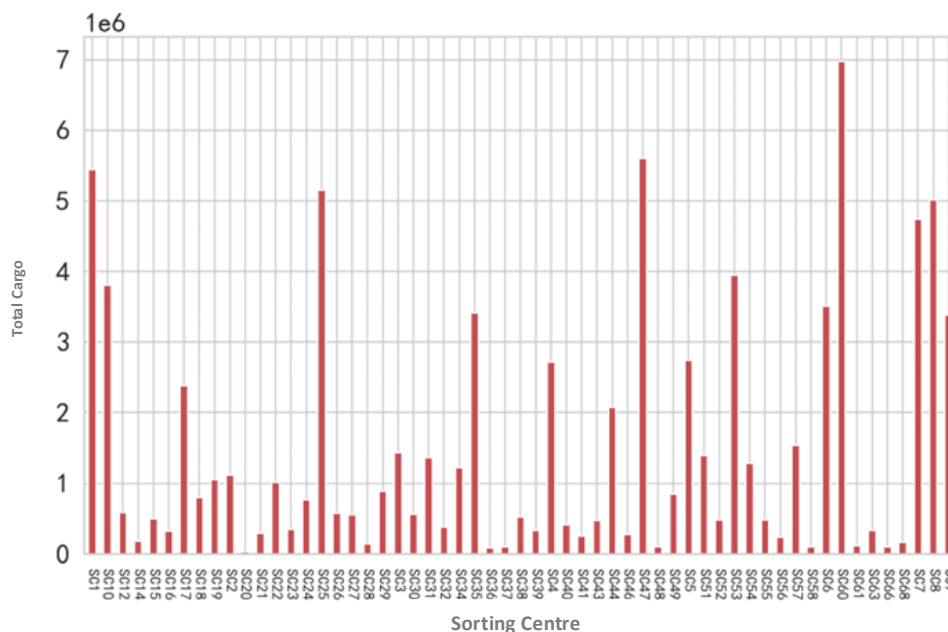
In terms of machine learning, Zhu Feilin et al [6] took Longyangxia Reservoir in Qinghai Province as an example and used Stacking fusion algorithm to establish an integrated prediction model, which used five heterogeneous prediction models such as ARMA, BP, LSTM, RF, and SVR to be fused, and used the hyper-parameter optimization method to determine the optimal parameters of each model. Li Chengju [7] selected 13 vegetation indices as input variables for the chlorophyll content

inversion model, and used Multiple Linear Regression, Support Vector Regression, Random Forest Regression, and Decision Tree Regression to construct a model for estimating the chlorophyll content of potato. Li Xiaoying [8] et al. selected six key monitoring stations in the Yangtze River estuary to analyse the response relationship between dissolved oxygen and other water quality factors. Three models, namely, improved support vector machine regression, artificial neural network and random forest, were used to predict the monthly average water quality data from 2004 to 2020 for comparative analysis. Zhuduan Dai [9] used three machine learning models, Random Forest Regression, Support Vector Machine Regression, and Multilayer Perceptron Regression, to build a dataset and perform training and prediction using UHPC raw material usage as a feature and 28-day compressive strength as a label. Wu Huizhen [10] used the meteorological observation data inside and outside the Venlo-type glass greenhouse of Nanjing University of Information Engineering to construct and validate the temperature prediction model for each season. Four algorithms, namely, multiple regression, BP artificial neural network, random forest and support vector machine, were used to construct and compare the seasonal forecasting models of daily mean temperature, daily minimum temperature and daily maximum temperature inside Venlo-type glass greenhouse to improve the forecasting accuracy.

The main challenge of the study is to accurately forecast cargo volumes for future time periods. Since changes in cargo volumes can be influenced by a variety of complex factors, including seasonal fluctuations, economic trends, holiday spending patterns, etc., this requires a high degree of flexibility and accuracy in the forecasting model. For this reason, the study uses machine learning models for forecasting. Compared with traditional time series models, machine learning models are more capable of handling non-linear relationships and capturing hidden patterns. By utilising a large amount of historical data, machine learning models can learn the complex patterns of changes in cargo volume and reveal the intrinsic connections between the data through feature engineering techniques. These features include the lagged values of historical shipments, the average value of the sliding window, etc., which help the model to better understand the time-series characteristics of the data. The study selected various models for comparative analysis such as Random Forest, Support Vector Machine and Neural Network. By evaluating their performance on the same dataset, the model that best suits the characteristics of the current data is selected. The study adopted a rolling prediction strategy, where the latest data is continuously incorporated into the model and the model is retrained to adapt to the new data environment. In this way, the timeliness and accuracy of the prediction results can be ensured.

## 2. Data preprocessing

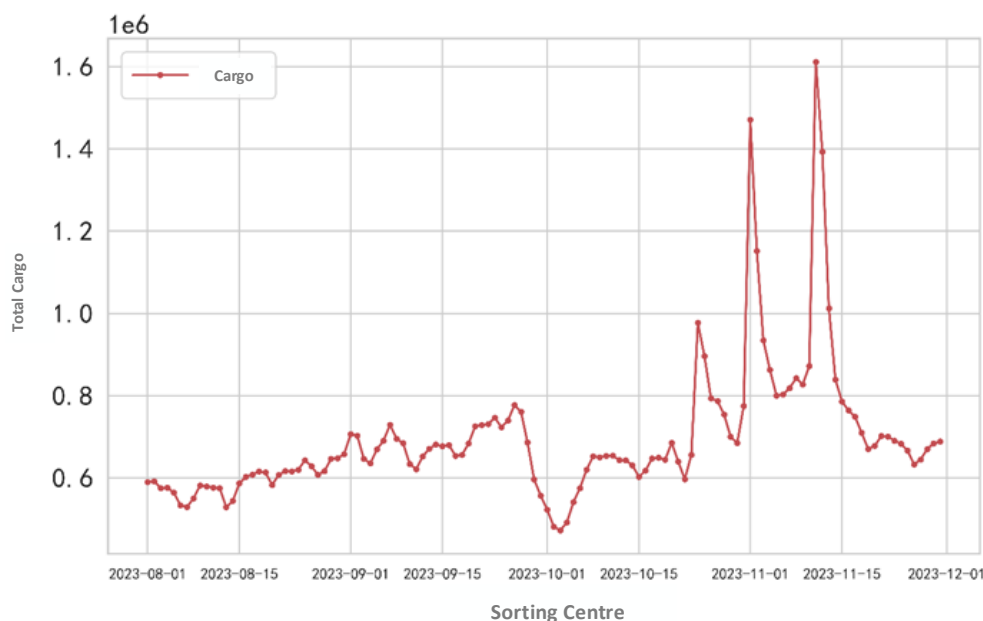
Data selected from the number of shipments per day at 57 sorting centres of a logistics company in the past 4 month. Since different sorting centres may be in different geographical locations, as well as there are differences in traffic, population and other elements, so to analyse the differences, the total cargo volume of different sorting centres is visualised, as shown in Figure 1.



**Figure 1.** Total volume of goods in different sorting centres

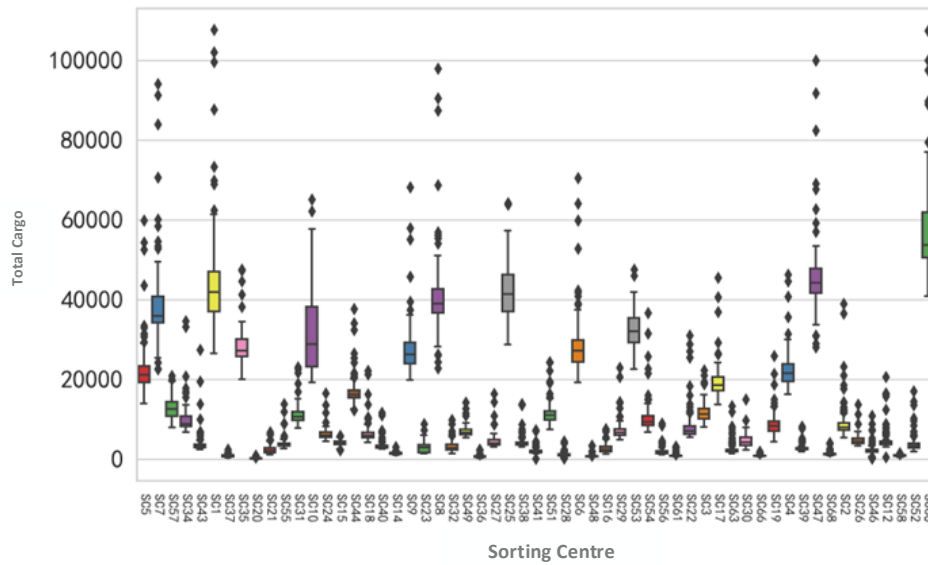
From the visualisation results, it can be seen that there are obvious differences in the total cargo volume of different sorting centres, with SC60 being the one with the highest total cargo volume, which is very different from the total cargo volume of other sorting centres, probably because different sorting centres have their own models, personnel and other factors, so when constructing the machine learning model in the follow-up, it is necessary to construct a separate model for each sorting centre separately.

After considering the differences in the volume of different sorting centres, the total volume trend of the sorting centre is visualised by date, to see the updated changes in the goods at a certain time period, as shown in Figure 2. There are obvious fluctuations in the total cargo volume change on different dates, and the change is obvious in November.



**Figure 2.** Total availability on different dates

Meanwhile, to further analyse the difference in the volume of goods in different sorting centres, a box-and-line diagram is visualised to analyse the statistical nature of different sorting centres, as shown in Figure 3.



**Figure 3.** Box line diagram

As can be seen in the box line plots for different sorting centres, there is also a significant difference in the mean and variance of the volumes of different sorting centres, which implies that the variance of a sorting centre with a high total volume of goods may not necessarily be large, further illustrating that there is a significant difference in the pattern of the volume trends of the individual sorting centres.

### 3. Random Forest, SVR model construction

Forecasting is enhanced by manually constructing a series of features. These features include cargo volumes over the past 12 days, 3-day and 6-day moving average cargo volumes, and time series-specific month, day, and weekday attributes. In addition, to capture the overall trend, also incorporated trend indicators to reflect the long-term pattern of cargo volume changes over time. Define the lagged value characteristics as:

$$Lag_i(t) = y_{t-i} \quad (1)$$

The three-day moving average MA3 (t) and the six-day moving average MA6 (t) are defined, respectively:

$$MA3(t) = \frac{1}{3} \sum_{k=0}^2 y_{t-1-k} \quad (2)$$

$$MA6(t) = \frac{1}{3} \sum_{k=0}^2 y_{t-1-k} \quad (3)$$

Given the differences in the characteristics of different sorting centres, these predictors need to be constructed individually for each sorting centre to ensure the accuracy and applicability of the model.

In terms of model selection, the study considered a variety of machine learning models and selected random forest, neural network, support vector machine and linear regression for modelling.

The idea of feature importance assessment with random forest is to see how much each feature contributes to each tree in the random forest, and then take the average value, and finally compare the contribution between features.

Where about the contribution is calculated as Gini index. The Gini coefficient is denoted by GI and is publicised as follows:

$$GI_m = \sum_{n=1}^n P_n(1 - P_n) \quad (4)$$

(Where:  $GI_m$  is the Gini coefficient of the data at node m and  $P_n$  represents the sample weights of the n categories.)

Then the importance of feature  $X_j$  at node  $m$ , i.e., the amount of change in the Gini index before and after branching at node  $m$ , is:

$$VIM_{jm}^{(Gini)} = GI_m - GI_1 - GI_r \quad (5)$$

(Where VIM is variable importance;  $GI_1$  and  $GI_r$  denote the Gini indices of two new nodes after branching, respectively.)

If the nodes of feature  $X_j$  appearing on a certain decision tree  $i$  in a random forest are set  $M$ , then the importance  $VIM_{ij}^{(Gini)}$  of feature  $X_j$  in the  $i$ th tree is:

$$VIM_{ij}^{(Gini)} = \sum_{m \in M} VIM_{jm}^{(Gini)} \quad (6)$$

Assuming that there are  $k$  trees in the random forest, the importance  $VIM_j^{(Gini)}$  of feature  $X_j$  is:

$$VIM_j^{(Gini)} = \sum_{i=1}^k VIM_{ij}^{(Gini)} \quad (7)$$

A non-linear estimation model is assumed when  $n$  input variables exist:

$$\hat{y} = f(x) = b + w^T \phi(x_i) \quad i \in \{1, 2, \dots, n\} \quad (8)$$

In the expression:  $b$  stands for bias;  $w^T$  is the weight vector;  $\phi(\cdot)$  is the high-dimensional feature space;  $x_i$  is the input vector.

SVR achieves the optimisation model by maximising the interval bandwidth with minimising the total loss, i.e., solving to obtain  $w^T$  and  $b$  in  $f(x)$ , where certain constraints need to be satisfied, viz:

$$\begin{cases} \min_{w,b} \frac{1}{2} \|w\|^2 \\ \text{s.t. } |y_i - (w^T \phi(x_i) + b)| \leq \varepsilon \quad i = 1, 2, \dots, n \end{cases} \quad (9)$$

In practice,  $\varepsilon$  is set too small to ensure that all sample points are in the interval bands, and  $\varepsilon$  is too large that the regression hyperplane will be biased by some anomalies, for this reason, SVR allows each sample  $(x_i, y_i)$  to add a relaxation variable  $\xi_i$  to avoid that it is not feasible to solve the optimisation problem. At this point the SVR objective function is expressed as:

$$\begin{cases} \min_{w,b,\xi^+,\xi^-} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i^+ + \xi_i^-) \\ -y_i + (w^T \phi(x_i) + b) \leq \varepsilon + \xi_i^+ \\ \text{s.t. } y_i - (w^T \phi(x_i) + b) \leq \varepsilon \quad i = 1, 2, \dots, n \\ \xi_i^+ + \xi_i^- \geq 0 \end{cases} \quad (10)$$

Where:  $\xi_i^+ + \xi_i^-$  denotes the upper and lower bound relaxation variables of the interval, respectively;  $C$  is the penalty coefficient, which is used to balance the complexity and loss of the model, too large a value of  $C$  will lead to the interval band is too narrow; smaller or tends to be close to 0 will make the interval band is too wide, which will affect the fitting effect, and this time, take the value of  $C=1$ .

In summary, this paper constructs a combined time series forecasting model based on Random Forest-SVR.

#### 4. Modelling Assessment

The accuracy and bias of the model predictions are checked by calculating the models' scores on the test set and analysing the residual plots Each model will be evaluated based on how well they fit the training set and how well they perform on the validation set. It is ensured that the final model

chosen will provide more reliable and accurate predictions of the volume of goods in the sorting centre.

Build and train separate prediction models for each sorting centre, as the total shipment volume of different sorting centres is affected by various subjective and objective factors. It is guaranteed that each model can accurately capture the unique shipment dynamics of each sorting centre.

The evaluation metrics used in this study are Mean Absolute Percentage Error (MAE), Mean Square Error (MSE), and Root Mean Squared Error (RMSE), which is the average of the absolute value of the deviation between all individual observations and the true value. The MAE is the average of the absolute values of all individual observations and the true value, which is less sensitive to outliers than the first two, and can reflect the actual situation of prediction error more realistically.

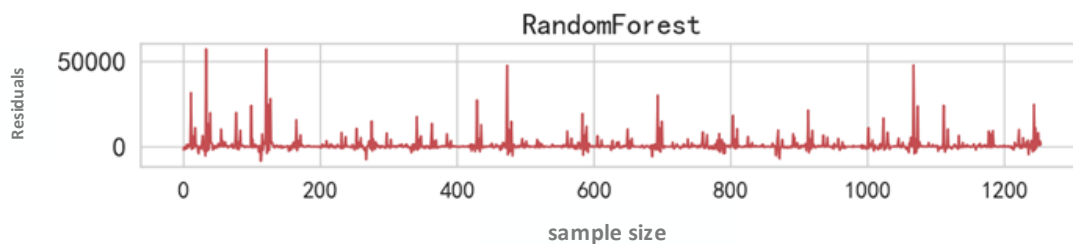
By training the model and evaluating the performance on the test set, some results are shown in Table 1.

**Table 1.** Average performance results for each algorithm

|      | RandomForest | SVR      | LinearRegression |
|------|--------------|----------|------------------|
| MAE  | 0.08256878   | 0.102536 | 0.0912297446     |
| MSE  | 0.02594959   | 0.028610 | 0.0291357415     |
| RMSE | 0.15717717   | 0.166337 | 0.1671507976     |

The results show that Random Forest is optimal on the test set.

After inverse normalising the data, for the results on the test set, the residuals of the random forest modes were plotted as shown in Figure 4.



**Figure 4.** Residual plots of the model

After that the predicted volume output for the next 30 days for each sorting centre is carried out using Random Forest.

For specific volume prediction, rolling prediction is used for prediction (lagged features need to be calculated for each day of data prediction), and data from the past 12 months are used as lagged value features to predict the volume of goods per day for the next 30 days, and some of the prediction results are shown in Table 2.

**Table 2.** Forecast of partial daily shipments for the next 30 days

| Sorting centre | Date      | Quantity  |
|----------------|-----------|-----------|
| SC5            | 2023/12/1 | 23709.3   |
| SC5            | 2023/12/2 | 23312.719 |
| SC5            | 2023/12/3 | 22400.319 |
| SC5            | 2023/12/4 | 22332.129 |
| SC5            | 2023/12/5 | 22245.019 |

Similarly, the construction of the random forest model is also carried out on the cargo hourly data, and the prediction of the hourly part of the cargo volume of each sorting centre for the next 30 days is increased on the basis of the original training data with hourly features, which is a more detailed prediction, and some of the prediction results are shown in Table 3.

**Table 3.** Hourly partial cargo forecast for the next 30 days

| Sorting centre | Date       | Hours | Quantity |
|----------------|------------|-------|----------|
| SC54           | 2023-12-01 | 0     | 193.59   |
| SC54           | 2023-12-01 | 1     | 163.06   |
| SC54           | 2023-12-01 | 2     | 152.14   |
| SC54           | 2023-12-01 | 3     | 138.67   |
| SC54           | 2023-12-01 | 4     | 129.26   |

## 5. Conclusion

This paper provides an in-depth discussion of machine learning-based logistics cargo forecasting techniques. To effectively deal with the complexity and variability of sorting centre cargo forecasting, the machine learning models of Random Forest and Support Vector Regression (SVR) are selected for in-depth study. In terms of features, lag features, trend features and cycle features are considered comprehensively, to construct a comprehensive and fine prediction model. Through systematic training and evaluation of various models, this paper finds that the random forest model shows excellent performance and stability in cargo volume prediction. Therefore, this paper finally chooses the random forest model as the core prediction tool to fit the prediction of the cargo volume of the sorting centre. To ensure the timeliness and accuracy of the prediction results, this paper further introduces a rolling prediction strategy. This strategy can dynamically update the model to capture new trends and changes in cargo flow in a timely manner, thus providing more accurate and real-time prediction data. In the future, this forecasting model will continue to be explored and optimised, and it is planned to introduce more advanced machine learning algorithms, such as deep learning models, to further improve the accuracy and efficiency of forecasting. In addition, consideration will be given to incorporating more external data, such as weather conditions and the impact of holidays, with a view to achieving accurate forecasts in a wider range of scenarios.

## References

- [1] Hu Jiaying. An introduction on the methods of cargo volume forecasting in express delivery industry [J]. Computer Knowledge and Technology, 2018, 14 (08): 152 - 153.
- [2] Zhang Panli. Research on optimisation of warehouse storage capacity of e-commerce FMCG front-end warehouse based on order prediction[D]. North China Electric Power University, 2020.
- [3] Antelope Liu. Research on full network routing simulation optimisation and algorithm for LTL logistics trunk network [D]. Shanghai jiaotong university, 2019.
- [4] Liu Di. Application of GWO-Elman model in port cargo throughput forecasting [J]. Water Transport Management, 2024, 46: (03).
- [5] YANG Yu-Ge, HAO Yang-Yang, WANG Yi-Wen. Port cargo throughput prediction based on NeuralProphet-LSTM combined model [J]. China Navigation, 2023, 46 (04): 85 - 92.
- [6] ZHU Feilin, CHEN Jiabi, ZHANG Mi, et al. An integrated approach to medium- and long-term runoff prediction based on multi-model fusion [J]. Hydropower, 2024, 50 (02): 6 - 13+29.
- [7] LI Chengju, LIU Yindu, QIN Tianyuan, et al. Estimation of chlorophyll content of potato leaves based on machine learning [J]. Spectroscopy and Spectral Analysis, 2024, 44 (04): 1117 - 1127.
- [8] LI Xiaoying, WANG Hua, WANG Yiqing, et al. Machine learning-based prediction model and evaluation of dissolved oxygen in the Yangtze River estuary[J/OL]. Environmental Science, 1 - 15 [2024-04-21].
- [9] Dai Zhuduan, Chen Jian, Zhou Xiaoyang. Research on UHPC compressive strength prediction based on machine learning regression model [J]. Concrete and Cement Products, 2023, (12): 23 - 27+33.
- [10] WU Huizhen, LI Dongsheng, YANG Zaiqiang, et al. Comparative analysis of seasonal forecasting models for Venlo-type glasshouse temperatures based on four algorithms [J]. China Agricultural Meteorology, 2024, 45 (02): 135 - 146.