Research On Cargo Volume Prediction at Sorting Centers Based on Multiple Machine Learning Models

JieBo Xing [1, *], YiFei Zhang [2]

[1] Leicester International Institute, Dalian University of Technology, Panjin, China, 124221

[2] School of Materials Science and Engineering, Inner Mongolia University of Science and Technology, Baotou, China, 014010

* Corresponding Author Email: xjb0522886@163.com

Abstract. Forecasting cargo volume is a crucial part of the workflow in sorting centers and has a significant impact on their operational capabilities. This paper considers a variety of machine learning methods to predict future cargo volumes at sorting centers. By analyzing the variance and mean differences in cargo volumes across different centers in the dataset, multiple features were extracted, including historical cargo data, time series features, and individual trend indicators for different centers. These features were used to train the feature sets for Random Forest models, Neural Network models, and Linear Regression models. Additionally, this paper evaluated the prediction performance of these machine learning models using the MAE, MSE, and RMSE metrics. Ultimately, the Random Forest model, which showed the best evaluation results, was selected to forecast future cargo volumes at sorting centers. The predicted results from the Random Forest model showed that for sorting center SC1, the MAE was 0.085, MSE was 0.313, and RMSE was 0.177; for sorting center SC35, the MAE was 0.072, MSE was 0.019, and RMSE was 0.176. These results indicate that the Random Forest model developed in this study is highly accurate in predicting cargo volumes at sorting centers.

Keywords: Comprehensive Evaluation of Multiple Machine Learning Models, Random Forest, Neural Networks, Linear Regression.

1. Introduction

In today's society, with the general increase in people's incomes, shopping demands have surged, leading to increasingly fierce competition in the logistics industry. In the logistics system, goods move from warehouses through sorting centers to various branches. Sorting centers, which segment packages according to different destinations, occupy a central role in the logistics network. Forecasting the cargo volume for sorting centers to improve sorting efficiency has become an essential part of operational management. Moreover, accurate cargo volume predictions can lay the groundwork for reasonable staff scheduling, facilitating the development of standardized, systematic, and innovative operational models at sorting centers, thus achieving significant development.

Li et al. [1] proposed a Quadratic Decomposition-Integration (SDE) method based on the Cuckoo Search Algorithm (CSA) for predicting air cargo volumes. They tested the method using cargo data from three different airports in China and concluded that it outperforms other baseline algorithms in terms of accuracy. Chung et al. [2] employed dynamics models based on system data to predict cargo volumes in North Korea. Hsu et al. [3] addressed the potential performance shortcomings of classical methods, such as time-series models, in scenarios with small and volatile datasets by introducing a hybrid forecasting model based on grey prediction techniques. D et al. [4] developed a novel DFA-ARIMAX (Dynamic Factor Analysis-ARIMAX Modeling) method for forecasting cargo volumes at the Port of Koper in the Adriatic Sea, and benchmark tests indicated that this method surpasses competing models. He et al. [5] developed an equal-dimensional Grey Markov model based on freight volumes at Chinese airports from 2010 to 2019, set against the backdrop of the pandemic. The findings reveal that this model outperforms the traditional GM (1,1) model in terms of forecasting accuracy and provides predictions that align more closely with actual cargo flows.

Data sources for this paper include datasets comprised of daily cargo volumes from 57 sorting centers over the past four months and hourly cargo volumes over the past 30 days from Dataset 1 and Dataset 2, respectively. (The following Table 1 shows the daily cargo volume graph for a specific sorting center from Dataset 1, and Table 2 shows the hourly cargo volume graph for a specific sorting center from Dataset 2).

Table. 1. Daily cargo volume at a specific sorting center

Sorting Center	Date	Cargo Volume
SC1	2023-11-16	51085
SC1	2023-9-10	39697
SC1	2023-8-12	33793
SC1	2023-8-15	34678
SC1	2023-8-10	34123
SC1	2023-11-5	56707
SC1	2023-10-13	41697
SC1	2023-10-8	41362
SC1	2023-8-27	36814
SC1	2023-11-25	44496

Table. 2. Hourly daily cargo volume at a specific sorting center

Sorting Center	Date	Hour	Cargo Volume
SC35	2023-11-30	20	2599
SC35	2023-11-8	16	139
SC35	2023-11-27	8	796
SC35	2023-11-18	14	174
SC35	2023-11-28	12	934
SC35	2023-11-21	4	1411
SC35	2023-11-21	13	99
SC35	2023-11-19	17	71
SC35	2023-11-9	1	1598
SC35	2023-11-4	15	197

This paper utilizes three machine learning models—Random Forest, Neural Network, and Linear Regression—to predict the cargo volume for the next 30 days and the hourly cargo volume for the next 30 days at sorting centers and evaluates the predictive results. By calculating and comparing the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) of the three models, it was found that the Random Forest model had the best performance. Consequently, this paper uses the Random Forest model to forecast both the daily and hourly cargo volumes of 57 sorting centers for the next 30 days. The workflow diagram for this study is shown in Figure 1 below.

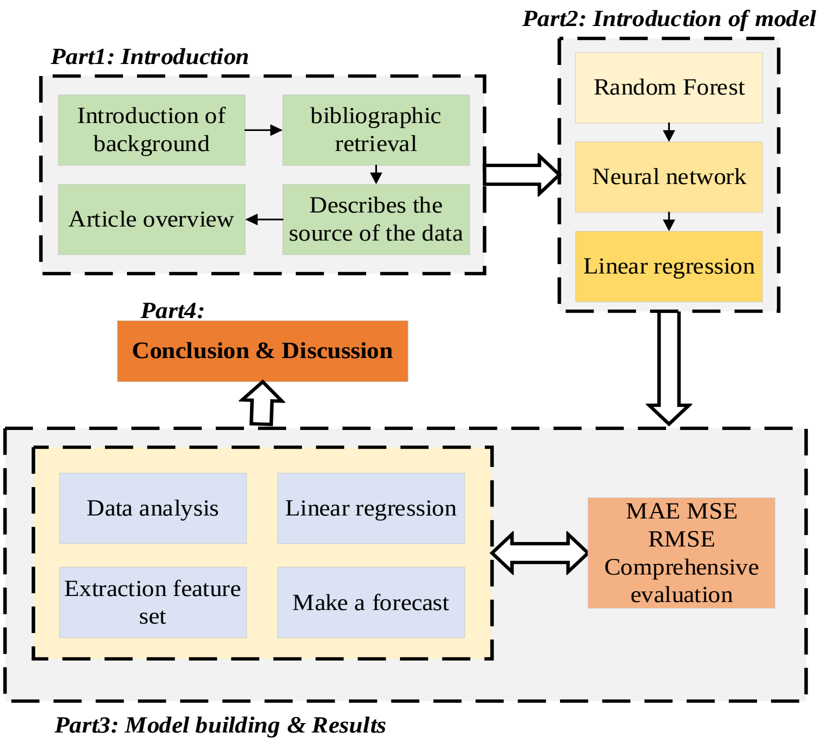

Figure 1. Workflow diagram

2. Model introduction

2.1. Fundamentals of the Random Forest model

The Random Forest model is a widely used machine learning algorithm that employs the outputs of multiple decision trees to arrive at a single result, as shown in Figure 2. Decision trees start with the most basic question and are then built up of a series of questions that form decision nodes of the tree. Each of these questions helps in determining the final answer. Here, observations that meet the criteria will follow the "yes" branch, while those that do not will follow an alternative branch. Decision trees use these questions to find the optimal method to train subsets and ultimately achieve the best results[6].

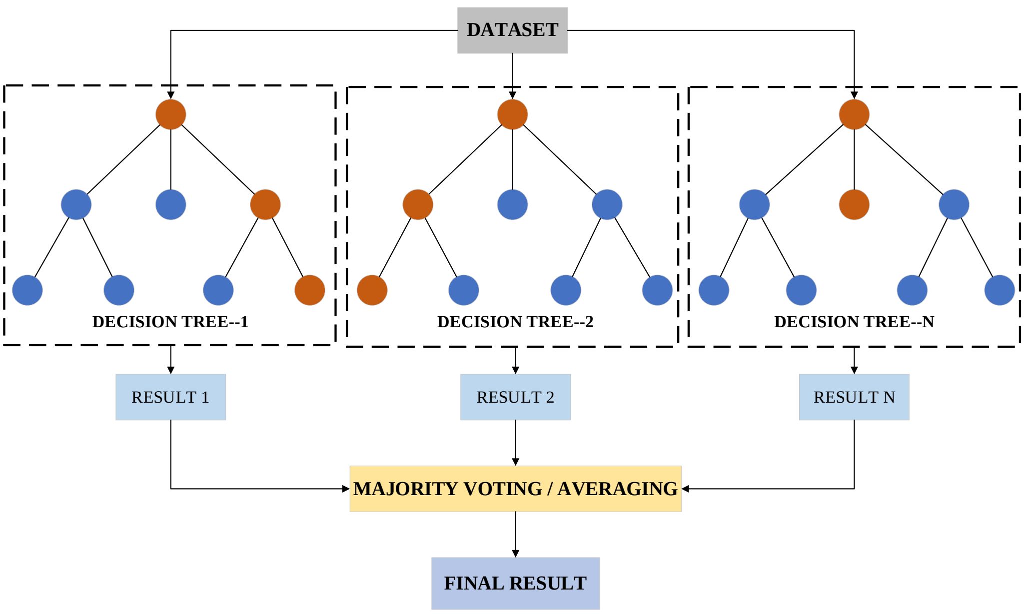

Figure 2. The working principle of the random forest model

When specifically used, the Random Forest model follows the following mathematical principles[7]:

$$F = S(|T_1(d_1), T_2(d_2), \cdots, T_n(d_n)|) \quad (1)$$

Where F is the final class, S is the selection function, Tn is the decision tree processing function, dn is the input data for the decision tree, and n is the number of decision trees. Based on these functions, the corresponding Random Forest prediction model can be established.

2.2. Fundamentals of the Neural Network model

Neural network models are typically defined as a type of machine learning model that recognizes phenomena by mimicking the collaborative functioning of neurons in the human brain. As shown in Figure 3, a neural network generally consists of an input layer, hidden layers, and an output layer, with interconnected nodes at each layer. Each node has a certain threshold; once this threshold is exceeded, the node is activated, transmitting data to the next layer. If not, the data does not pass to the next layer. Neural networks also improve their accuracy through training, making them a powerful tool in the field of artificial intelligence[8].

In practical use, there could assume that an artificial neuron receives n inputs: [x1,x2,⋯,xn], with corresponding weights [w1,w2,⋯,wn] assigned to each input value. Next, the output can be obtained by multiplying each input value by its corresponding weight and adding a bias term b[9]:

$$z = (\sum_{i=1}^{n} x_i \times w_i) + b \quad (2)$$

However, the above formula is only applicable to linear models and does not suit most real-life situations. Therefore, when building nonlinear models, we may consider using the following activation functions[9]:

$$output = a(z) = a[(\sum_{i=1}^{n} x_i \times w_i) + b] \quad (3)$$

By doing so, each neuron can use its respective activation function to achieve activation from input to output, thus establishing the neural network model.

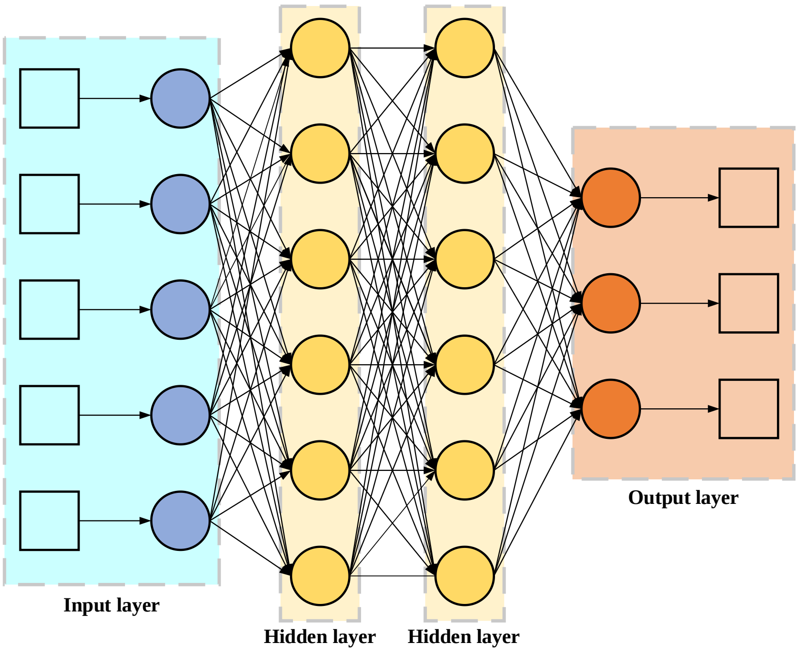

Figure 3. Working principle of neural networks

2.3. Fundamentals of the Linear Regression model

Linear regression is commonly used to quantify and predict the relationship between predictor variables and their corresponding predicted values. This is achieved by fitting a regression equation to the dataset, which is then used for future predictions[10].

3. Results

3.1. Data analysis and feature extraction

Step 1: Upon analyzing the dataset, this study identifies that different sorting centers exhibit distinctive characteristics due to their varied geographical locations. In Figure 4, this paper has illustrated the total cargo volumes of each sorting center, revealing significant disparities in cargo volumes among them. Consequently, when considering the development of machine learning models, it is essential to account for the unique circumstances of each sorting center to facilitate more detailed modeling.

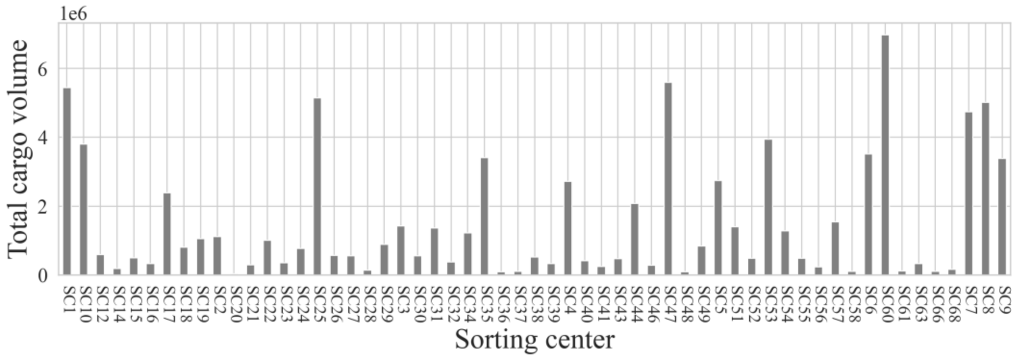

Figure 4. Total cargo volume across sorting centers

Step 2: Additionally, the total daily cargo volume at each sorting center varies. In Figure 5, this paper has depicted various trends in the total cargo volumes over different periods, often reflecting specific underlying circumstances. For example, a significant increase in cargo volumes around November, as shown in Figure 5, may be associated with the "Double Eleven" promotional events prevalent in real life. Utilizing these volatile data points in model training could enhance machine performance and improve the accuracy of predictive outcomes.

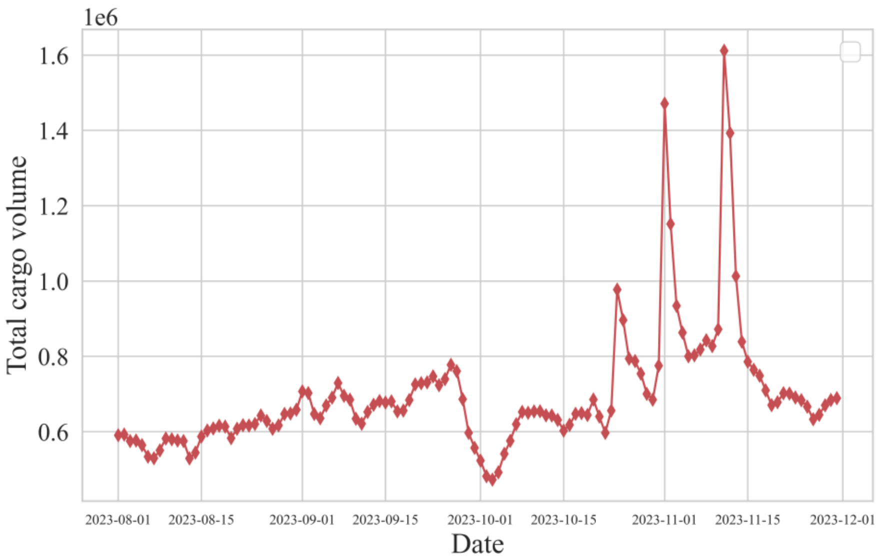

Figure 5. Cargo volume by date

Step 3: Subsequently, this paper analyzed the distribution of variance, mean, and other statistical measures for each sorting center and produced box plots as shown in Figure 6. Comparing these with Figure 4, it could observe that high cargo volume at a sorting center does not necessarily correlate with high variance, indicating that factors other than total cargo volume, such as variance and mean, also significantly influence the outcomes. Similarly, this paper employed these statistics in training our models to achieve more precise predictions

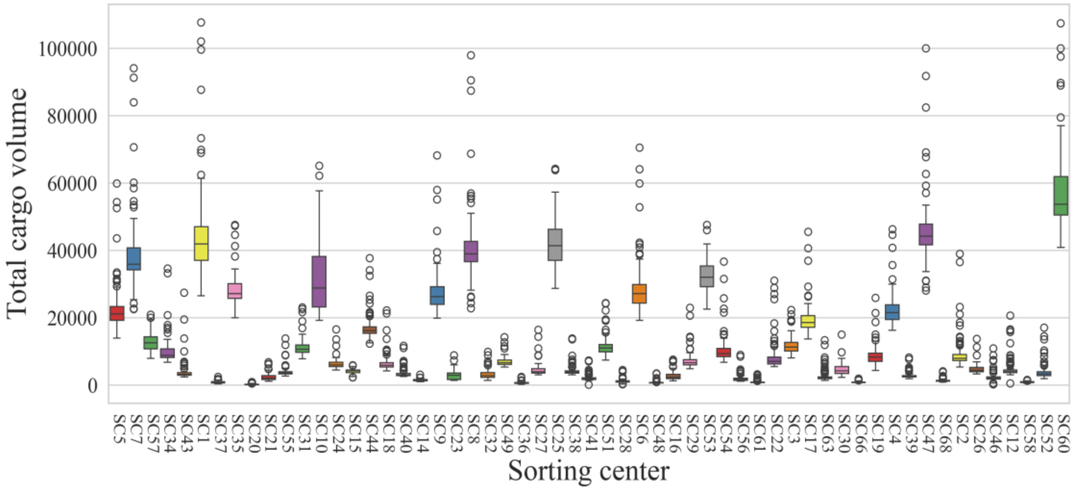

Figure 6. Mean and variance of total cargo volume across sorting centers

Step 4: Furthermore, the discrepancies in total cargo volume across different sorting centers may also relate to their specific attributes. For instance, some centers might primarily handle daily necessities, thus making them less susceptible to the impacts of certain activities compared to other centers.

Step 5: After analyzing the features of Dataset 1, this paper shifted our attention to Dataset 2, which had undergone linear interpolation. As Dataset 2 is compiled from hourly cargo statistics over 30 days for each sorting center, it might differ from Dataset 1. Therefore, this paper initially compared the two datasets to check for consistent trends. Detailed comparison is illustrated in Figure 7.

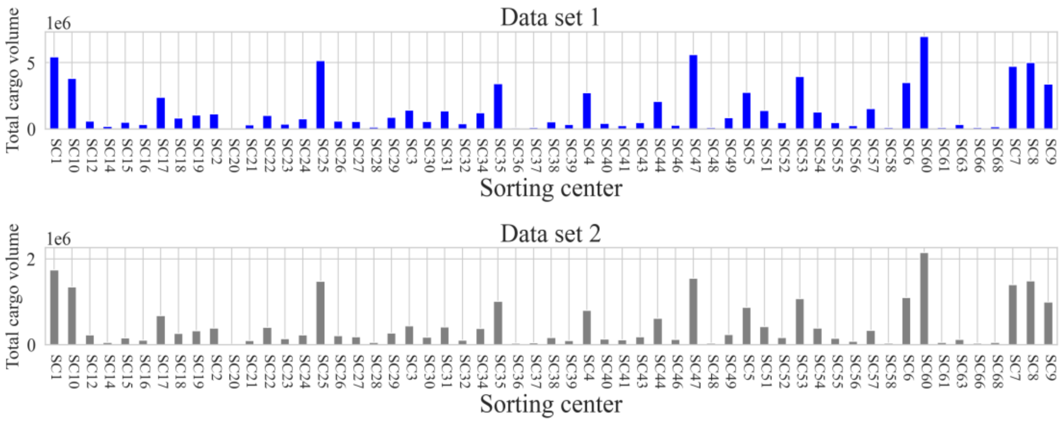

Figure 7. Comparison of appendices one and two

As observed from Figure 7, sorting centers with high daily cargo volumes also exhibit high hourly volumes, demonstrating highly similar distribution profiles. The primary difference between the two lies in the total volume of cargo. Thus, the machine learning approaches applied to Dataset 1 are generally applicable to Dataset 2 as well.

3.2. Model development process

In this case, to avoid the error due to the tendency of ARIMA models to stabilize over long prediction intervals, this paper opted for the Random Forest model for prediction.

(1) First, this paper manually constructed several features based on the dataset, including historical cargo data, time series features (such as months), and various trend indicators for different sorting centers.

(2) Next, this paper initially forecast the next 30 days. Here, in addition to the Random Forest model, this paper incorporated other machine learning models including Neural Networks and Linear Regression models. This paper evaluated each model based on how well they fit their training sets and performed on validation sets. Then, there could selecte the most accurate model with the smallest error by comparing their scores on the test set and analyzing residual plots.

(3) Given the differences among the three models, this paper trained them separately and evaluated each using the sorting metrics MAE (Mean Absolute Error), MSE (Mean Squared Error), and RMSE (Root Mean Squared Error). The specific formulas are cited in[11],[12],[13].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{4}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{6}$$

Herein, n represents the number of samples, $\hat{y}_i$ denotes the predicted values, and yi refers to the actual values.

This article selected four sorting centers for evaluation. The assessment results for these four centers are displayed in Table 3 below.

Table. 3. Partial assessment results

Sorting Center	Evaluation Metrics	Random Forest	Neural Network	Linear Regression
	MAE	0.052	0.065	0.046
SC10	MSE	0.007	0.016	0.009
	RMSE	0.086	0.119	0.088
	MAE	0.073	0.129	0.088
SC24	MSE	0.026	0.031	0.029
	RMSE	0.161	0.179	0.167
	MAE	0.136	0.158	0.143
SC15	MSE	0.028	0.031	0.037
	RMSE	0.169	0.199	0.189
	MAE	0.079	0.089	0.079
SC44	MSE	0.028	0.025	0.029
	RMSE	0.168	0.168	0.169

Based on Table 3, it is evident that the predictions from the Random Forest model are more accurate compared to the other two models.

Simultaneously, by comparing these predictions with real outcomes, there could have constructed the residual plot shown in Figure 8 below.

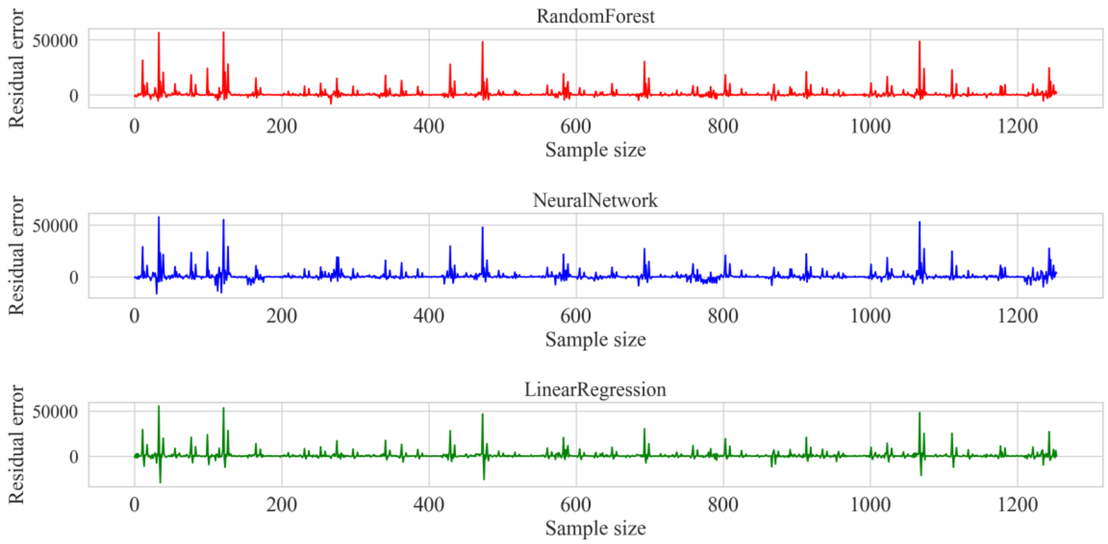

Figure 8. Residual plots for the three machine learning models

From the residual plot mentioned above, it can be observed that the Random Forest model maintains a high level of accuracy. Therefore, this paper has chosen to use the Random Forest model to predict the future cargo volumes for the sorting centers over the next 30 days as well as on an hourly basis.

3.3. Prediction results

Following the data extraction and result evaluation, this paper presents partial results from the predictions of future cargo volumes using the Random Forest model.

Table 4 shows predicted cargo volumes for a randomly selected sorting center over the next 30 days (selected data for 10 days), and Table 5 forecasts hourly cargo volumes for the same center over the next 30 days (selected data for 10 hours).

Table. 4. Predicted results for the next 30 days

Sorting Center	Date	Cargo Volume
SC1	2023-11-21	45994
SC1	2023-11-22	45352
SC1	2023-11-23	45169
SC1	2023-11-24	44719
SC1	2023-11-25	44496
SC1	2023-11-26	42879
SC1	2023-11-27	41085
SC1	2023-11-28	44200
SC1	2023-11-29	45217
SC1	2023-11-30	45707

Table. 5. Hourly predictions for the next 30 days

Sorting Center	Date	Hour	Cargo Volume
SC35	2023-12-18	1	1823.56
SC35	2023-12-18	2	1466.63
SC35	2023-12-18	3	1400.729
SC35	2023-12-18	4	1351.57
SC35	2023-12-18	5	1422.88
SC35	2023-12-18	6	31.579
SC35	2023-12-18	7	1.11
SC35	2023-12-18	8	973.249
SC35	2023-12-18	9	1535.8
SC35	2023-12-18	10	1458.679

The achieved prediction values using the Random Forest model are as follows: for sorting center SC1 the MAE is 0.085, MSE is 0.313, and RMSE is 0.177, for SC35 the MAE is 0.072, MSE is 0.019, and RMSE is 0.176. These results underscore the high precision of the Random Forest model.

4. Discussions

4.1. Model Improvements

The model presented here is an idealized construct based on the topics and data provided. It might deviate significantly from real-world results in the event of unforeseen circumstances. Hence, this paper proposes several potential directions for improving the model. With more comprehensive and complete information, the model could be more closely aligned with real-world conditions:

(1) Weather Factors: In this paper modeling, the impact of weather on outcomes was minimized. However, in real life, it is essential to include weather conditions when training the model. For example, monsoon regions experience slippery roads during the rainy season, and high-latitude areas might have roads covered with snow in the winter. Therefore, incorporating weather forecasts for the upcoming month could allow for a more realistic prediction of cargo volume.

(2) Technological Factors: In more developed areas, sorting centers often undergo rapid technological changes. The cargo volume in the next month might differ significantly from that in the current month due to these advances, so using shorter spans of data for these centers could yield more accurate forecasts.

4.2. Model Applications

The model developed for predicting future cargo volumes at sorting centers has practical applications, as outlined below:

(1) Similar Work Environments: This prediction model can also be applied to scenarios such as patient volume forecasting in hospitals. Just as with cargo volumes, patient numbers typically fluctuate due to seasonal viruses, exhibiting trends similar to those in our model. By combining different hospitals' historical patient data, time-series features of patient numbers, and trends, it is possible to predict patient volumes and thus assist hospitals in operating more effectively.

(2) Staff Scheduling at Sorting Centers: Based on the predictions from this model, staff scheduling at sorting centers can be optimized. By allocating more staff during peak cargo periods, sorting centers can enhance efficiency and prevent the accumulation of cargo, leading to smoother operations.

5. Conclusion

Throughout this discussion, this paper proposes selecting various metrics such as cargo volume across different sorting centers, the monthly fluctuations in cargo volume at these centers, and the mean and variance of cargo volume. Additionally, this paper manually constructed features including historical cargo data of sorting centers, time series characteristics, and individual cargo volume trend indicators for each sorting center. These were used to train the Random Forest model, the Neural Network model, and the Linear Regression model. A comparison of the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) across these models revealed that the Random Forest model achieved the highest prediction accuracy. Consequently, this paper employed the Random Forest model to forecast future cargo volumes for sorting centers.

Beyond predicting cargo volumes, sorting centers can also implement reforms and innovation tailored to their specific circumstances. For instance, high-volume centers could consider introducing advanced automated sorting systems and establishing intelligent storage systems. In conjunction with the cargo volume prediction model, these technologies could enhance sorting efficiency and boost service capabilities. In today's diverse competitive landscape, the battle ultimately boils down to a contest of technology and innovation. Staying ahead in technological prowess can distinctly position a competitor to excel in the marketplace.

In summary, the development of sorting centers should keep pace with modern advancements, innovating and optimizing in areas such as human resources, operations, and technology. This involves not only enhancing operational efficiency but also focusing on employee welfare and environmental sustainability, ultimately aiming for sustainable development and providing superior logistics services for the well-being of the public.

References

[1] Li H, Bai J, Chen X, et al. A new secondary decomposition-ensemble approach with cuckoo search optimization for air cargo forecasting[J]. Applied Soft Computing, 2020, 90: 106161.

[2] Chung T, Jeon J. System dynamics-based prediction of North Korean port volumes[J]. Asian Journal of Shipping and Logistics, 2021, 37(4): 337-344.

[3] Hsu C I, Huang Y C, Wong K. A Grey hybrid model with industry share for the forecasting of cargo volumes and dynamic industrial changes[J]. Transportation Letters, 2018, 12(1): 25-36.

[4] Dragan D, Keshavarzsaleh A, Intihar M, et al. Throughput forecasting of different types of cargo in the Adriatic seaport Koper[J]. Maritime Policy and Management, 2020, 48(1): 19-45.

[5] He H, Ding H, Liu S, et al. Airport cargo volume forecasting based on equal dimensional and new information grey Markov model [C]//Lecture Notes in Computer Science. 2022: 264-276.

[6] Meltzer R. What is Random Forest? [Beginner's Guide + Examples] [EB/OL]. CareerFoundry, 2023-08-31[2024-05-06]. URL: https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/

[7] Random Forest — the science of Machine learning & AI[EB/OL]. The Science of Machine Learning & AI, [n.d.] [2024-05-06]. URL:https://www.ml-science.com/random-forest

[8] Terra J. What is a Neural Network? - Caltech [EB/OL]. Caltech, 2023-05-06. URL: https://pg-p.ctme.caltech.edu/blog/ai-ml/what-is-a-neural-network

[9] Cristina S. Calculus in action: Neural networks[J]. Machine Learning Mastery, 2022, March 15.

[10] Bobbitt Z. How to Make Predictions with Linear Regression[J]. Statology, 2021, July 27.

[11] Tim. Absolute Error & Mean Absolute Error (MAE) - Statistics How to[J]. Statistics How To, 2020, December 28.

[12] Szczepanek A. MSE calculator. Mean Squared Error[J]. Mean Squared Error, 2024, January 18.

[13] RMSE: Root Mean Square Error - Statistics How To[J]. Statistics How To, 2024, March 29.