WEP Warwick Evans Publishing

Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

International Air Cargo Volume Forecasting based on the Stacking Fusion Model

Lei Tang [1,#], Yining Zhou [2,#], Peilan Zhao [3,#,*]

[1] College of Science, Civil Aviation University of China, Tianjin, China, 300300

[2] School of Business, Macau University of Science and Technology, Macau, China, 999078

[3] College of Air Traffic Management, Civil Aviation Flight University of China, Deyang, China, 618300

* Corresponding author: 17757896568@163.com

[#] These authors contributed equally.

Abstract: Accurate forecasting of international air cargo volumes is crucial for air cargo carriers, which optimizes operational efficiency and profitability and strengthens market competitiveness and risk response capabilities. This study focuses on the open-source cargo data of over 100 international airlines in India from January 2015 to March 2017. Firstly, the missing data must be dealt with. Then, histograms and probability density plots are used to depict the distribution of the cargo volume and analyze the dynamics of the cargo volume of each company at annual and quarterly levels. Subsequently, key factors affecting cargo volumes are analyzed, with a special focus on the cargo performance of the Top 5 airlines, which is visualized in box plots. Given the time-series nature of cargo data, this study adopts the data sequential partitioning method. It innovatively utilizes the Stacking ensemble learning framework, integrating regression models such as Random Forest, Decision Tree, Extreme Gradient Boosting, and K-Nearest Neighbor as the base learners. The meta-learner is constructed to integrate the outputs of each base model, and the hyperparameters of the model are optimized by grid search and cross-validation. Comparison results show that the Stacking model has a higher R^2 value and lower MSE than any single base learners, indicating that it is more effective in international air cargo volume prediction. The prediction model based on stacking ensemble learning provides an effective strategy for improving air cargo prediction accuracy.

Keywords: Air Cargo, Stacking, Random Forest Regressor, Decision Tree Regressor, K-Nearest Neighbor Regressor.

1. Introduction

As an important part of global trade, the air cargo industry is poised for new opportunities as global trade increases and e-commerce grows, yet industry participants need to respond to changing market conditions and challenges. In order to improve efficiency and reduce costs, it is vital for air cargo carriers to improve the accuracy of air traffic forecasts.

To solve this problem, researchers have begun to explore combining or fusing the prediction results of multiple single learners with certain strategies to predict the international air passenger or cargo volume problem. For example, Wenjun et al. used a gray GM model and a regression analysis model to combine them to predict air cargo volume using a combined weighted average [1]. Zhu et al. used quadratic exponential smoothing and a gray forecast model to establish a single forecasting model. Then, the weights of the single forecasting model are determined by taking the inverse of the variance [2]. The composite pattern is simple and easy to perform, but focuses on integrating and optimising existing predictions, and cannot improve the performance of a single prediction model on its own.

Unlike the composite pattern, the Stacking fusion model can reduce the error by further predicting the result of the base learners. Liu Bo et al. proposed a model fusion method based on XGBoost and Stacking to predict the bus load, using different XGBoost as the basic learner that can weaken the effect of a single meta-model error and the PSO algorithm to optimize the system parameters [3]. Wang Fei et al. proposed a machine learning model based on Stacking multi-model fusion to realize the life prediction of IGBTs, which fused four complementary machine learning algorithm models and greatly reduced the mean square error of prediction [4]. Liu Dong-xu et al. screened as the base model of the Stacking model by cross-validation of the training set training, resulting in a significant improvement in prediction accuracy compared to a single machine learning model [5].

The model fusion based on Stacking can fully utilize the advantages of multiple base learners to improve the prediction performance. Therefore, this paper innovatively considers the application of this learning method to the prediction of international air passenger or cargo volume to further improve the prediction accuracy. In this paper, processing the data with missing values is required, then exploring the data, building the air cargo volume prediction model based on Stacking, and performing grid search and cross-validation hyperparameter optimization to get the final air cargo volume prediction model. The structure of the paper is shown in Figure 1 below.

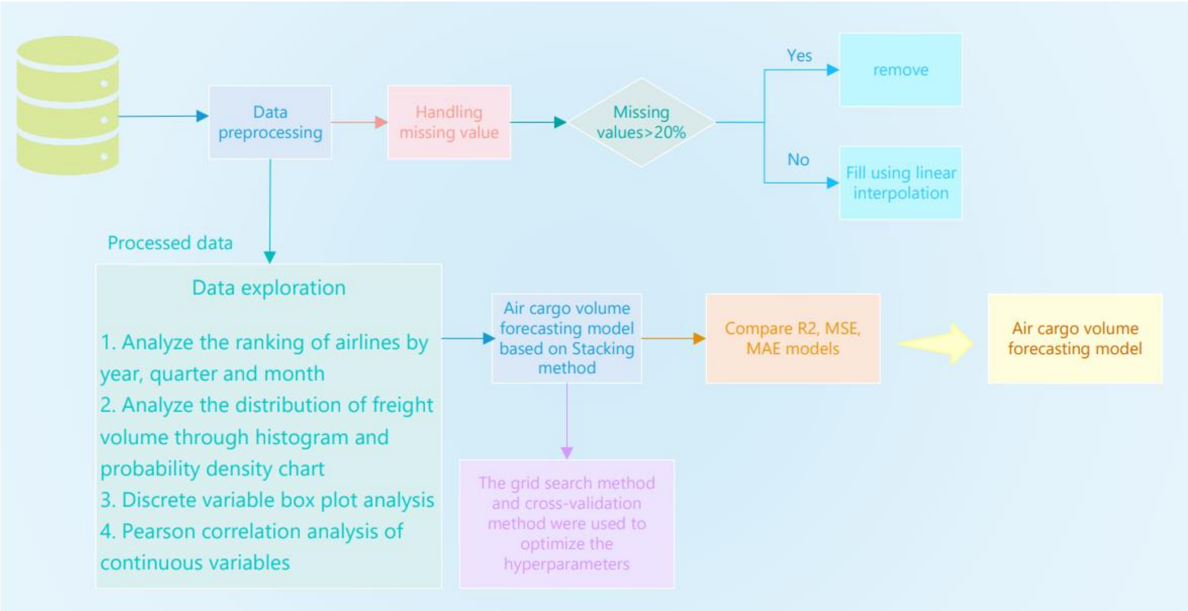

Figure 1. Structure of the paper

2. Data Exploration

2.1. Data Source and Cleaning

The data in this article is supported by https://www.kaggle.com/, an open-source data site. This data set is the cargo data of more than 100 Indian international airlines from January 2015 to March 2017. It contains the following data dimensions: YEAR, MONTH, QUARTER, ALRLINE NAME, CARRIER TYPE, FREIGHT TO INDIA, and FREIGHT FROM INDIA, with a total of 2334 pieces. Some of the data is shown in Table 1 below.

Table 1. Raw data

YEAR	MONTH	QUARTER	AIRLINE NAME	CARRIER TYPE	PASSENGERS TO INDIA	PASSENGERS FROM INDIA	FREIGHT TO INDIA	FREIGHT FROM INDIA
2015	JAN	Q1	AIR INDIA	DOMESTIC	258876.00	274220.00	3320.63	4186.30
2015	JAN	Q1	AIR INDIA EXPRESS	DOMESTIC	95581.00	116600.00	0.00	0.00
....	....	....	....	....	....	....	....	....
2017	MAR	Q1	YEMENIA AIRWAYS	FOREIGN	622.00	715.00	0.00	6.30

This dataset revealed the presence of missing values. Therefore, the following missing value processing strategy is used for processing, grouping statistics according to each airline; if the percentage of missing data in the group is more than 20%, the airline data will be deleted directly; the rest of the missing percentage is less than 20%, the interpolation method will be used to fill in.

2.2. Cargo volume distribution

Air cargo volume is the sum of outbound cargo volume and inbound cargo volume. Cargo volume can comprehensively reflect the overall scale and task volume of air transportation enterprises in cargo transportation. Therefore, in this paper, 'FREIGHT FROM INDIA' and 'FREIGHT TO INDIA' are summed up to get the total cargo volume.

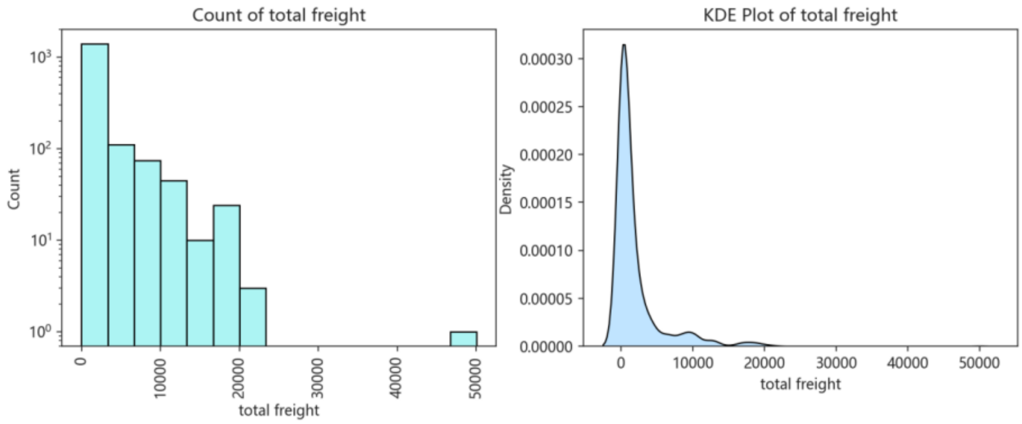

Figure 2. Histogram and Probability Density Plot of Freight Volume Distribution

Observation through the Figure 2 shows that most of the airlines' total transportation volume is concentrated in 0-3000, which shows the phenomenon of positively skewed distribution and long-tailed companies with larger values. Similarly, the probability density plot also shows that the peak highest point is 0-3000, indicating that the total transportation volume is mostly concentrated here.

2.3. Distribution of cargo volume by different airlines

2.3.1. Annual data

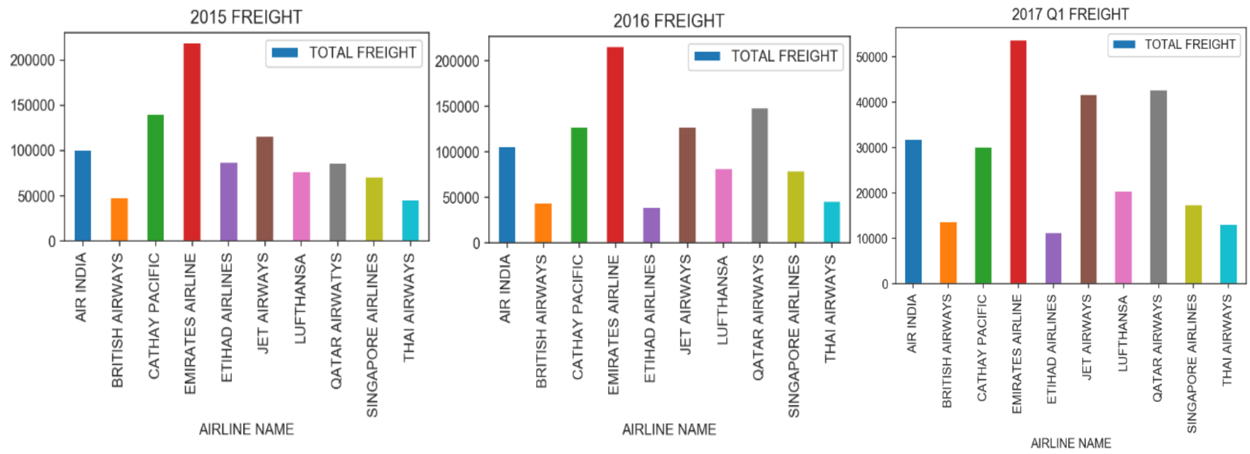

Figure 3. Distribution of Cargo Volume of Top 10 Airlines in Different Years, 2015-2017

Summing the cargo volume of each airline separately in terms of a year, the total cargo volume of each company in 2015, 2016, and 2017 was calculated, and the top 10 airlines are shown in Figure 3 above.

Observing the image, it is found that EMIRATES AIRLINE has been number one in terms of air traffic for all three years and is far ahead of the second place. As can be seen in Table 2, in 2015, the first to tenth place in terms of traffic are EMIRATES AIRLINE, CATHAY PACIFIC, JET AIRWAYS, AIR INDIA, ETIHAD AIRLINES, QATAR AIRWATYS, LUFTHANSA, SINGAPORE AIRLINES, BRITISH AIRWAYS, THAI AIRWAYS.

Table 2. Top 10 carriers in terms of total traffic in 2015

RANK	AIRLINE NAME	TOTAL FREIGHT(t)
1	EMIRATES AIRLINE	218889.70
2	CATHAY PACIFIC	140524.80
3	JET AIRWAYS	115932.30
4	AIR INDIA	101145.80
5	ETIHAD AIRLINES	87536.65
6	QATAR AIRWATYS	86573.15
7	LUFTHANSA	77281.24
8	SINGAPORE AIRLINES	70812.11
9	BRITISH AIRWAYS	48587.79
10	THAI AIRWAYS	45771.11

2.3.2. Quarterly data

Each airline's cargo volume was summed up quarterly, and each company's total cargo volume was calculated separately for the four quarters of the period 2015-2017. It was observed that the number one cargo volume for all four quarters was EMIRATES AIRLINE from Figure 4.

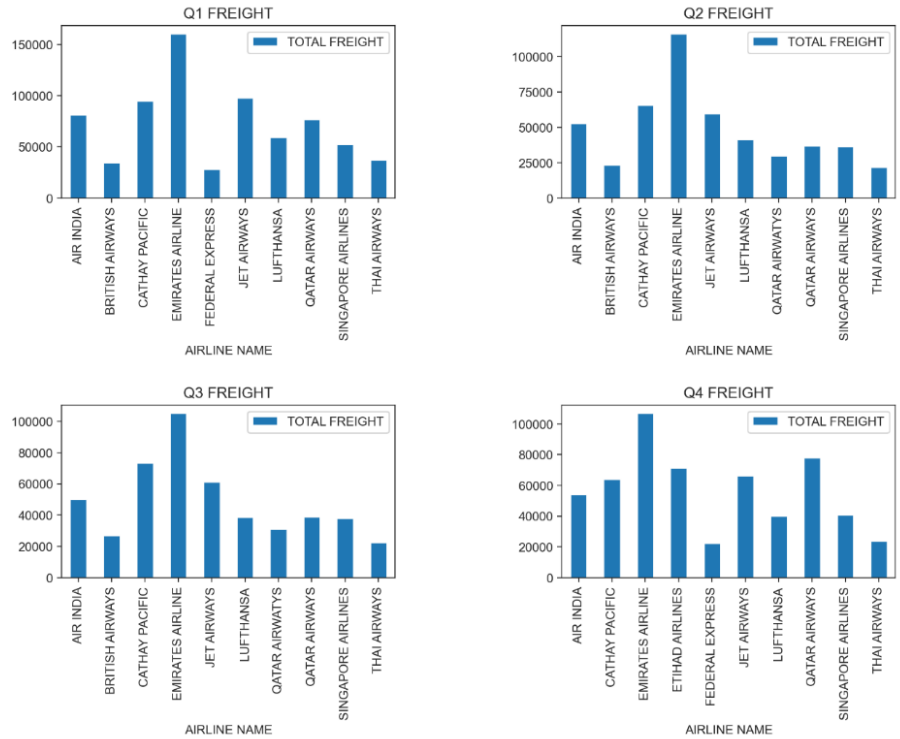

Figure 4. Transportation Volumes by Quarter

In the first quarter traffic ranking, the first to tenth-ranked airlines in terms of traffic are EMIRATES AIRLINE, JET AIRWAYS, CATHAY PACIFIC, AIR INDIA, QATAR AIRWAYS, LUFTHANSA, SINGAPORE AIRLINES, THAI AIRWAYS, BRITISH AIRWAYS, and FEDERAL EXPRESS from Table 3.

Table 3. Top 10 carriers by volume in Q1

RANK	AIRLINE NAME	Q1 FREIGHT(t)
1	EMIRATES AIRLINE	160299
2	JET AIRWAYS	97741
3	CATHAY PACIFIC	94708
4	AIR INDIA	81421
5	QATAR AIRWAYS	76867
6	LUFTHANSA	59449
7	SINGAPORE AIRLINES	52247
8	THAI AIRWAYS	36972
9	BRITISH AIRWAYS	34603
10	FEDERAL EXPRESS	28403

In Q2 and Q3, the rankings of some companies changed. This shows that the ranking of airlines in terms of volume fluctuates significantly when calculated on a quarterly basis.

2.3.3. Comparison of outbound and inbound cargo volumes

As shown in Figure 5, the volume of 'FREIGHT FROM INDIA' is greater than the volume of 'FREIGHT TO INDIA' when calculated on a quarterly basis and when calculated on a monthly basis.

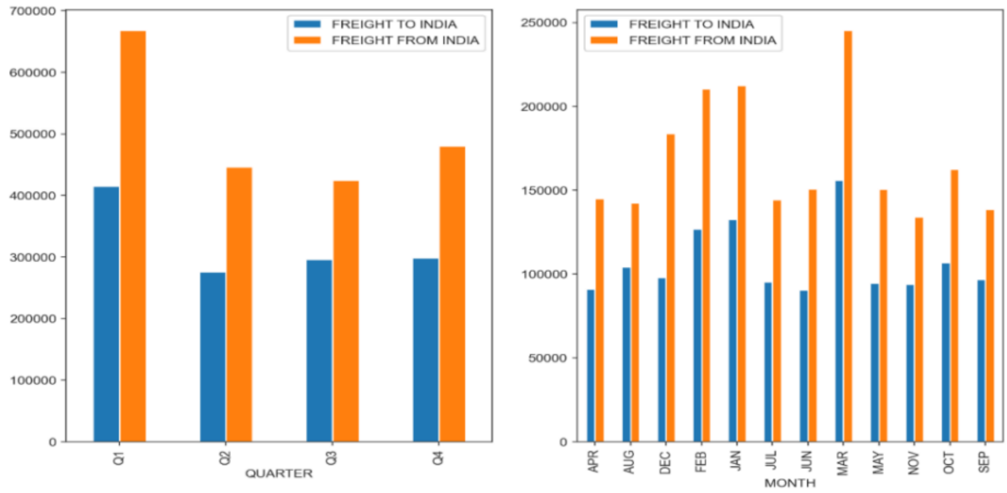

Figure 5. Quarterly and monthly data on airline cargo volumes

As can be seen in Table 4, in the standard deviation calculation, the total standard deviation, quarterly standard deviation, and monthly standard deviation of 'FREIGHT FROM INDIA' are greater than those of 'FREIGHT TO INDIA'.

Table 4. The standard deviation of FREIGHT TO INDIA and FREIGHT FROM INDIA

Std	FREIGHT TO INDIA	FREIGHT FROM INDIA
Total Standard Deviation	1460.42	2468.28
Quarterly Standard Deviation	62528.92	109225.83
Monthly Standard Deviation	20198.30	35772.49

2.3.4. Calculation of linear correlation

As shown in Figure 6 below, the highly positively correlated variables with 'TOTAL FREIGHT' are 'FREIGHT FROM INDIA' and 'FREIGHT TO INDIA'. Meanwhile, 'YEAR', 'QUARTER', and 'MONTH' correlate poorly with 'TOTAL FREIGHT'.

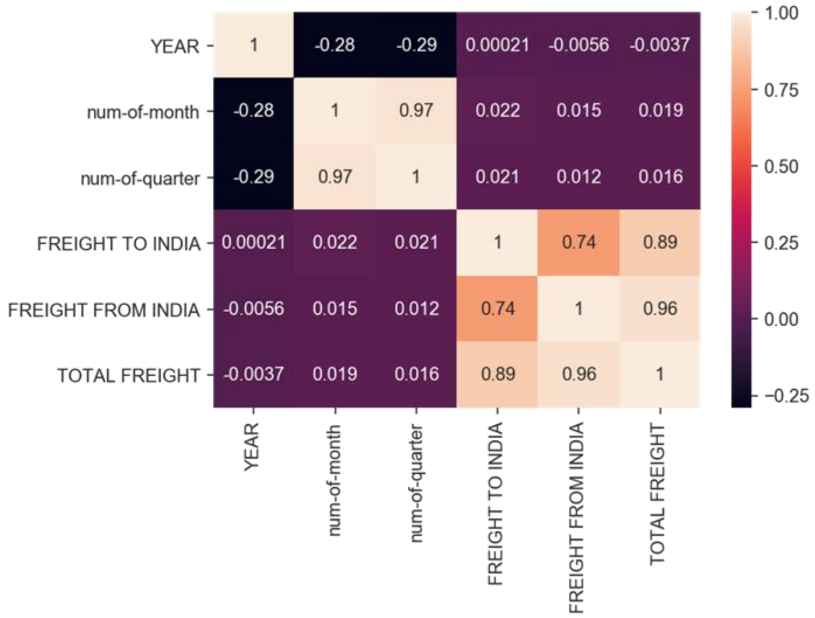

Figure 6. Heat map of correlation analysis of freight volume factors

2.3.5. Separate calculation according to airlines

In this paper, the collected data is grouped according to the airlines and counted 'TOTAL FREIGHT' and do the descending order to get the top five airlines as EMIRATES AIRLINE, CATHAY PACIFIC, JET AIRWAYS, AIR INDIA, QATAR AIRWAYS.

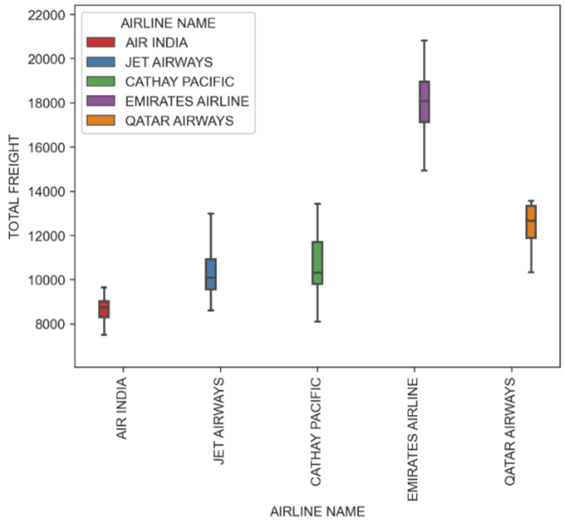

Figure 7. Top5 Airlines Cargo Volume Box Plot

From Figure 7, the width of the box is wider for the shipments of two airlines, EMIRATES AIRLINE, and CATHAY PACIFIC, indicating that the shipments of these two companies are more discrete relative to the other companies. While AIR INDIA has a narrower box width, means its shipments are less discrete, and most of the sales volume is less than the median. The median of the shipments of EMIRATES AIRLINE and QATAR AIRWAYS is located in the middle of the box, which indicates that the distribution of the shipments is roughly normal. CATHAY PACIFIC and JET AIRWAYS have their median located in the lower half of the box, indicating that they both have a right-skewed distribution of traffic.

3. Modeling establishment

3.1. Data segmentation

Correct data division can ensure that the model can effectively predict future time points in practical applications. The research data in this paper is time series data, due to the time-dependent nature of time series data, so when dividing the data, it is necessary to ensure that the data points in the training set are earlier than those in the test set. The data from 2015 to October 2016 is used as the training set for model training; the data from November 2016 to March 2017 is used as the test set for evaluating the generalization ability of the model.

3.2. Overview of Modeling Algorithm and Establishment

Decision tree is a method used for data classification and regression analysis that consists of nodes and edges. Nodes are categorized into internal and leaf nodes, where internal nodes represent a feature or attribute of the data, and leaf nodes represent a category [6]. The random forest model is an ensemble learning approach that trains multiple decision trees and aggregates their predictions. Each tree is independent, allowing for parallel processing during training, which enhances the model's speed [7]. XGBoost is a highly efficient ensemble learning algorithm that enhances classification by combining multiple weak learners. It surpasses other GB algorithms with system and algorithmic improvements, delivering superior prediction performance in less time [8].The basic principle of the KNN algorithm is to select the K data points with the closest distance as their nearest neighbors by calculating the distance between the data to be predicted and each data point in the training set [9]. Then, the data values to be predicted are based on the values of these K nearest neighbors.

In this paper, we use Stacking integrated learning to train the Random Forest Regressor, Decision Tree Regressor, XGBoost Regressor, and KNN Regressor as the base model and then construct a meta learner to be trained with the output of the base model as input to get the final prediction results [10]. The modeling flow is shown in Figure 8 below.

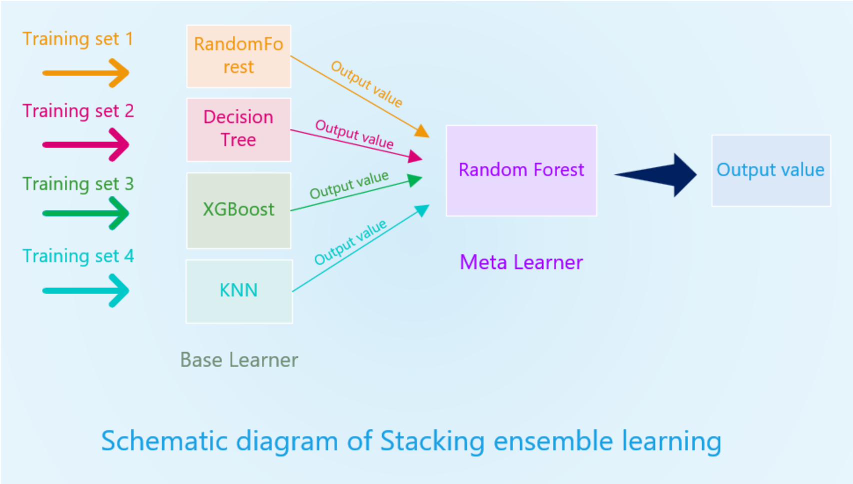

Figure 8. Structure of the model

In this paper, the optimal hyperparameters for the stacking model are determined through grid search and cross-validation. Grid search tries every possibility among all candidate parameter choices by loop traversal. It evaluates the performance of each combination using the k-fold cross-validation method and finally finds the best hyperparameters, as shown in the Table 5 below.

Table 5. Hyperparameters of Stacking Regressor model

level	Model	Hyperparameter
Base learners	Random Forest Regressor	n_estimators=30 min_samples_leaf=10 random_state=0
	Decision Tree Regressor	min_samples_leaf=7 random_state=0
	XGB Regressor	xgbregressor__gamma=0 xgbregressor__learning_rate=0.40 xgbregressor__max_depth=2 xgbregressor__n_estimators=100
	KNeighbors Regressor	n_neighbors = 3
Meta learner	Random Forest Regressor	n_estimators=500 min_samples_leaf=10 random_state=0

3.3. Model Evaluation

Model evaluation metrics are essential for assessing performance, with R^2, MSE, and MAE being key in regression. R^2, though straightforward, can be deceptive with large datasets, as it may appear high despite poor data explanation. MSE, sensitive to significant errors, highlights model deviations but is heavily swayed by outliers. In contrast, MAE, less outlier-prone than MSE, offers intuitive results but may downplay minor errors. These metrics collectively provide a nuanced assessment of model performance.

Table 6. Stacking Regressor Model Evaluation

Level	Model	R^2	MSE	MAE
Base learners	Random Forest Regressor	0.95	629908.41	348.40
	Decision Tree Regressor	0.93	935437.90	386.01
	XGB Regressor	0.95	658272.44	436.84
	KNeighbors Regressor	0.54	5983154.36	1249.47
Meta learner	Random Forest Regressor	0.96	578916.71	336.69

From the above Table 6, it can be seen that the MSE of single learners such as Random Forest Regressor, Decision Tree Regressor, XGBoost Regressor, KNeighbors Regressor is larger; Stacking model has the largest R^2 and the smallest MSE, which gives the optimal fitting results. Therefore, the stacking model is chosen for the prediction of international air cargo volume.

3.4. Prediction results

The accuracy of predicting international air cargo volume using Stacking ensemble learning method is shown below in Figure 9.

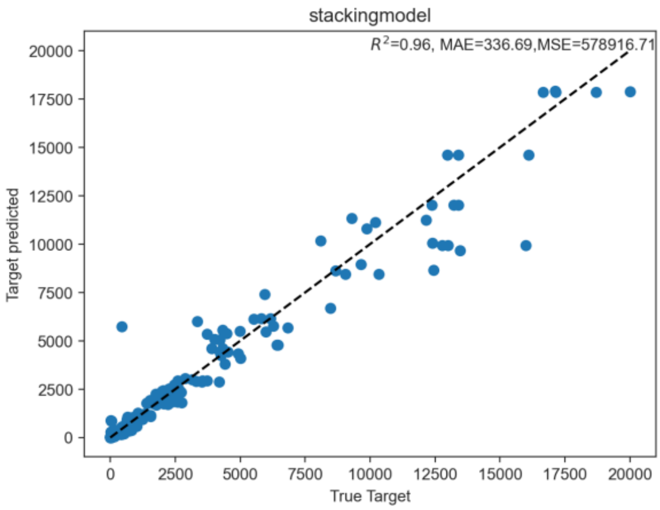

Figure 9. Scatterplot of predicted and true values of the Stacking model

It can be seen that the real and predicted values are uniformly distributed around the 45-degree diagonal line, which means that the prediction error of the model is small. The error is uniformly distributed within the range of each prediction value, without the situation that the error is larger in some areas and smaller in other areas.

4. Conclusions

This article innovatively considers the application of Stacking fusion model to forecast international cargo traffic to enhance the forecasting accuracy further. The study is based on the cargo data of more than 100 global airlines in India, and it is found that the cargo volume carried by different airlines varies considerably through box plots, while the historical cargo data is analyzed at multiple time spans, such as annually and quarterly. Through the Pearson correlation coefficient to determine the relationship between the influencing factors and the cargo volume, it is found that the two present a nonlinear relationship, so the subsequent use of nonlinear models for modeling. Eventually, this paper adopts Stacking integrated learning method, which integrates models such as Random Forest, Decision Tree, XGBoost, KNeighbors, etc., and obtains the best prediction model by optimizing hyperparameters. The evaluation shows that the Stacking fusion model is better than the single machine learning model in terms of accuracy, and the use of this model in the prediction of international air cargo volume can further improve the accuracy of the prediction of international air cargo volume to help airlines optimize the allocation of resources to improve the efficiency of resource utilization, and the accurate prediction of the cargo volume makes them respond to the market changes faster so as to enhance the competitiveness of the market. It also helps the airline companies to formulate more scientific and reasonable strategic goals.

In this paper, when exploring the factors affecting air freight traffic, it mainly focuses on the dimensions of time and airlines, and fails to comprehensively cover all the potential influencing factors, such as bad weather conditions leading to flight delays or cancellations, which in turn affects the cargo volume; the level of economic development in different regions exerts a specific influence on the demand for air freight; in addition, policies and regulations are one of the most important factors influencing the volume of air freight traffic. In summary, in order to further improve the model prediction accuracy, more data on weather, regional economic differences, industrial demand, and policies and regulations need to be collected and analyzed. By taking these factors into account, the trend of air freight traffic can be more comprehensively grasped.

References

[1] WEN Jun, JIANG You-hui, FANG Wen-qing. Optimal combination forecasting model o-f air cargo volume[J]. Computer Engineering and Applications,2010,46(15):215-217.

[2] ZHU Zhi-yu, LIU Yan. Combination Forecast of Air Cargo Volume in China Based on Time Sequence Model[J]. Journal of Xi'an Aeronautical Unive-rsity,2017,35(05):65-70.

[3] LIU Bo, QIN Chuan, JU Ping, et al. Short-term bus load forecasting based on XGBoos-t and Stacking model fusion[J]. Electric Power Automation Equipment,2020,40(3):147-153.

[4] WANG Fei, HUANG Tao, YANG Ye. Study on Machine Learning Algorithms for Life Prediction of IGBT Devices Based on Stacking Multi-model Fusion[J]. COMPUTER SC-IENCE,2022,49(6A):784-789.

[5] LIU Dong-xu, LI Mingming, SHAO Lei, ZOU Zong-shu. Prediction of end-point carbo-n content and temperature in AOD based on stacking model fusion[J/OL]. Steelmaking:1-12[2024-06-24].

[6] WANG Ying. Application of decision tree model in soybean futures price trend forecast[D]. Northwest Normal University,2021.002135.

[7] ZHANG Rui-jie, ZHOU Chun-yan, CHEN Hui, et al. Inversion of $PM_{2.5}$ in the Beijing-Tianjin-Hebei region using a random forest model based on GF-5B satellite[J/OL]. Chi-na Environmental Science,1-12[2024-08-03].

[8] Peng Bai-xue, Chen Qing-hua, Ji Jia-dong. Fault analysis of refrigeration system based on XGBoost and SHAP[J]. Cryo & supercond,2024,52(07):89-96.

[9] CHU Xin-yuan, LU Ai-zhen, ZHANG Jing-xin. A study on the prediction o-f crude oil price based on KNN model[J]. Monthly Price Magazine,2021, (05):15-22.

[10] SHANG Juan-ye. Remote resource transmission load prediction based on Stacking ense-mble learning[J]. information technology,2024, (06):94-99+104.