WEP Warwick Evans Publishing

Transactions on Economics, Business and Management Research
ISSN: 2960-1789, eISSN: 2960-2254 | Volume 11
EMFT 2024

Research on Logistics Cost Prediction Based on Random Forest Regression Modeling

Shuya Zhao *

School Of Economics And Management, Xi'An Technological University, Xi 'An, China, 710021

* Corresponding Author Email: zhaoshuya2024@163.com

Abstract. With the rapid development of globalization and e-commerce, logistics industry has become a key link between production and consumption. However, the efficient management and control of logistics costs has become an important challenge for enterprises to enhance their competitiveness. The scientific control of logistics cost depends on accurate cost prediction. In this paper, we first process the missing values of open source logistics cost data, and analyze the logistics cost and numerical influencing factors as skewed distribution with the help of distribution probability density plot and histogram, so we use logarithmic transformation to eliminate the skewness. The box plot and quartile method are used to visualize and process the outliers of discrete factors. For the cleaned data, the chi-square test and correlation coefficient method are applied to screen important features and remove redundant information, respectively. Then the random forest regression model was constructed to predict logistics costs, and the model parameters were optimized by grid search and cross-validation. Finally, the importance of the features of the prediction model is also ranked, and the model prediction mechanism is analyzed in depth to provide a scientific basis for logistics cost control.

Keywords: Lasso Regression, Random Forest, Logistics Cost Forecasting.

1. Introduction

With the rapid development of China's economy and the deepening of globalization, the logistics industry, as an important link between production and consumption, has become increasingly important. However, while the logistics industry is developing rapidly, it is also facing challenges such as difficult cost control and low operational efficiency. The volatility and uncertainty of logistics costs not only increase the business risks of enterprises, but also restrict the overall development level of the logistics industry. Therefore, how to predict logistics costs scientifically and accurately has become the key for logistics enterprises to enhance competitiveness and realize sustainable development.

Wang Lijuan et al [1] proposed a PCA-EBP neural network model for the problems of information overlap between the factors affecting the cost of coal logistics as well as the poor prediction results of BP neural network for multi-noise samples and small sample problems. The results are better than radial basis neural network (RBF) and support vector regression machine (SVR). Sun Changfeng et al. [2] take the logistics cost statistics from 2009 to 2013 as the basis, use the gray correlation model, the least squares linear regression model and the time series model to effectively detect the logistics cost, and analyze the prediction results. The results show that the prediction error of the gray correlation model is significantly smaller than that of the other two models. Feng Yanqiao et al [3] introduced BP neural network to train the learning samples of ship logistics and distribution cost, fit the changing characteristics of ship logistics and distribution cost, so as to realize the prediction of ship logistics and distribution cost, and the results show that the prediction accuracy of ship logistics and distribution cost of BP neural network is not only higher than that of the comparative methods by more than 5% on average, but also the stability is better.

Currently, the prediction methods of logistics cost mainly include empirical method, statistical method and machine learning method. The empirical method relies on the subjective judgment of experts and industry experience, but often lacks accuracy and repeatability; the statistical method

predicts logistics costs by establishing mathematical models [4], but the model complexity is high and it is difficult to adapt to the complex and changeable logistics environment; and the machine learning method, especially integrated learning algorithms such as Random Forest Regression, with its powerful data processing capabilities and nonlinear modeling capabilities, has shown a broad application prospect in the field of logistics cost prediction [5]. It shows a broad application prospect [5]. Although the random forest regression algorithm shows good performance in logistics-related forecasting, there are still some shortcomings in the previous research. For example, the missing values and outliers in logistics data are not sufficiently handled, which leads to a decrease in the model prediction accuracy; in terms of feature selection, most of the researches adopt methods based on experience or statistics, and lack a systematic feature screening machine. Therefore, this paper further explores the logistics cost prediction method based on random forest regression algorithm on the basis of previous studies. A more perfect data preprocessing process is constructed, the deskewing of data is added, a more scientific feature screening method is adopted, and the interpretability of the model is carried out at the same time. It aims to improve the accuracy and stability of logistics cost prediction and provide more effective support for cost control and operation optimization of logistics enterprises. It also provides strong theoretical support and practical guidance for the digital transformation and intelligent upgrading of the logistics industry.

In this paper, the data are first processed with outliers, and then discrete variables and continuous variables are screened for features using chi-square test and regression correlation, respectively, to establish a random forest regression logistics cost prediction model, grid search and cross-validation are used to find the model's optimal hyper-parameters, the optimized random forest regression algorithm is applied to predict the logistics costs, and finally, the importance of the features in the prediction of the logistics costs is ranked to explain the model. The technology roadmap of this paper is shown in Figure 1 below.

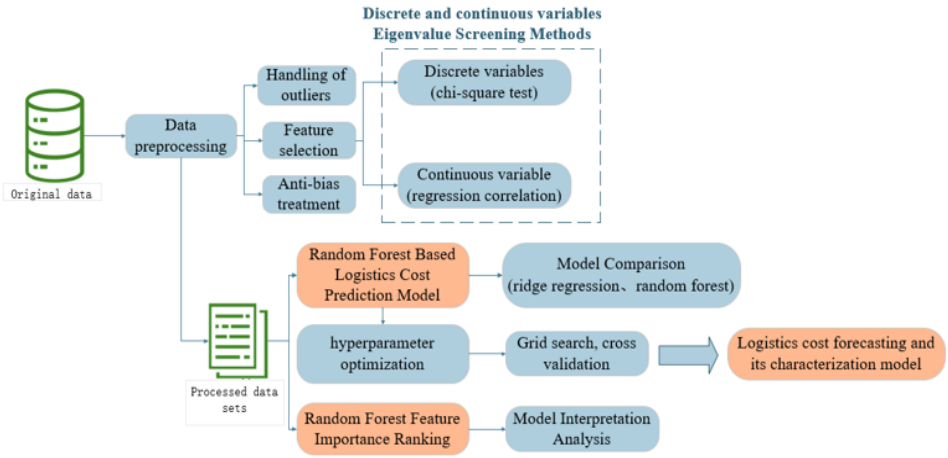

Figure 1. Technology roadmap.

2. Data Exploration

2.1. Introduction to Logistics Data

In the logistics industry, cost prediction based on logistics data is a key part of improving the operational efficiency and market competitiveness of enterprises. In this paper, we access to open source datasets from https://www.kaggle.com and use it as a base dataset for data analysis and exploration. The public dataset has a total of 10,324 data records, recording detailed information from the supplier procurement, warehousing, distribution to customer service and other aspects, including the status of the goods, transportation routes, time, trade conditions, mode of transportation and so on, a total of 33 key data attributes, and some of the original data are shown in Table 1 below. There are missing values and abnormal values in the above data, which need to be pre-processed, otherwise it is easy to cause deviation in logistics cost prediction. At the same time, the data contains many

types of variables, such as numerical variables and discrete variables, and different types of variables require different processing methods.

Table 1. Partial presentation of raw data.

ID	PO / SO #	ASN/DN #	Country	...	Dosage	Dosage Form	Line Item Quantity	Line Item Value	Pack Price
1	SCMS-4	ASN-8	Côte d'Ivoire	...	N/A	Test kit	19	551	29
3	SCMS-13	ASN-85	Vietnam	...	10mg/ml	Oral suspension	1000	6200	6.2
...	...	...	...	...	...	...	...	....	...
23	SCMS-87	ASN-57	Nigeria	...	10mg/ml	Oral solution	416	2225.6	5.35
44	SCMS-139	ASN-130	Zambia	...	200mg	Capsule	135	4374	32.4

2.2. Data Cleaning

2.2.1. Analysis of missing values

Missing values are a common problem in data analysis and machine learning projects and can be caused by various reasons such as omission during data collection, equipment failure, data entry errors, etc. In this study, missing values were analyzed and it was found that transportation mode, dose, and program insurance ($) had missing values that needed to be subsequently addressed to ensure data quality. There are no missing values in the remaining columns

2.2.2. Treatment of numerical variables

The following variables are numeric, such as: ID, Unit of Measure (Per Pack), Line Item Quantity, Line Item Value, Pack Price, Unit Price, Line Item Insurance (USD), and by looking at the raw data, we found that Weight (Kilograms), Freight Cost (USD), should be numeric columns, but the code did not return them when filtering, so we looked at these two columns and found that there are strings. Kilograms), Freight Cost (USD), should be numerical columns, but the code does not return when filtering, so check the two columns, found that there are strings. Need to carry out the corresponding data cleaning operations.

Replaces all Freight Included in Commodity Cost values in the Freight Cost column with 0, indicating that freight is included in the commodity cost. Convert the values in the Weight (Kilograms) column to a numeric type; values that cannot be converted are replaced with NaN. similarly, convert the values in the Freight Cost (USD) column to a numeric type. Use the average of the Weight (Kilograms) column to populate all NaN values in that column. Use the average of the Freight Cost (USD) column to populate all NaN values in that column.

2.2.3. Discrete variable handling

The following variables in this data are partially discrete, and the sample sizes for each category in each variable are shown below in Figure 2 as (a)-(d), respectively.

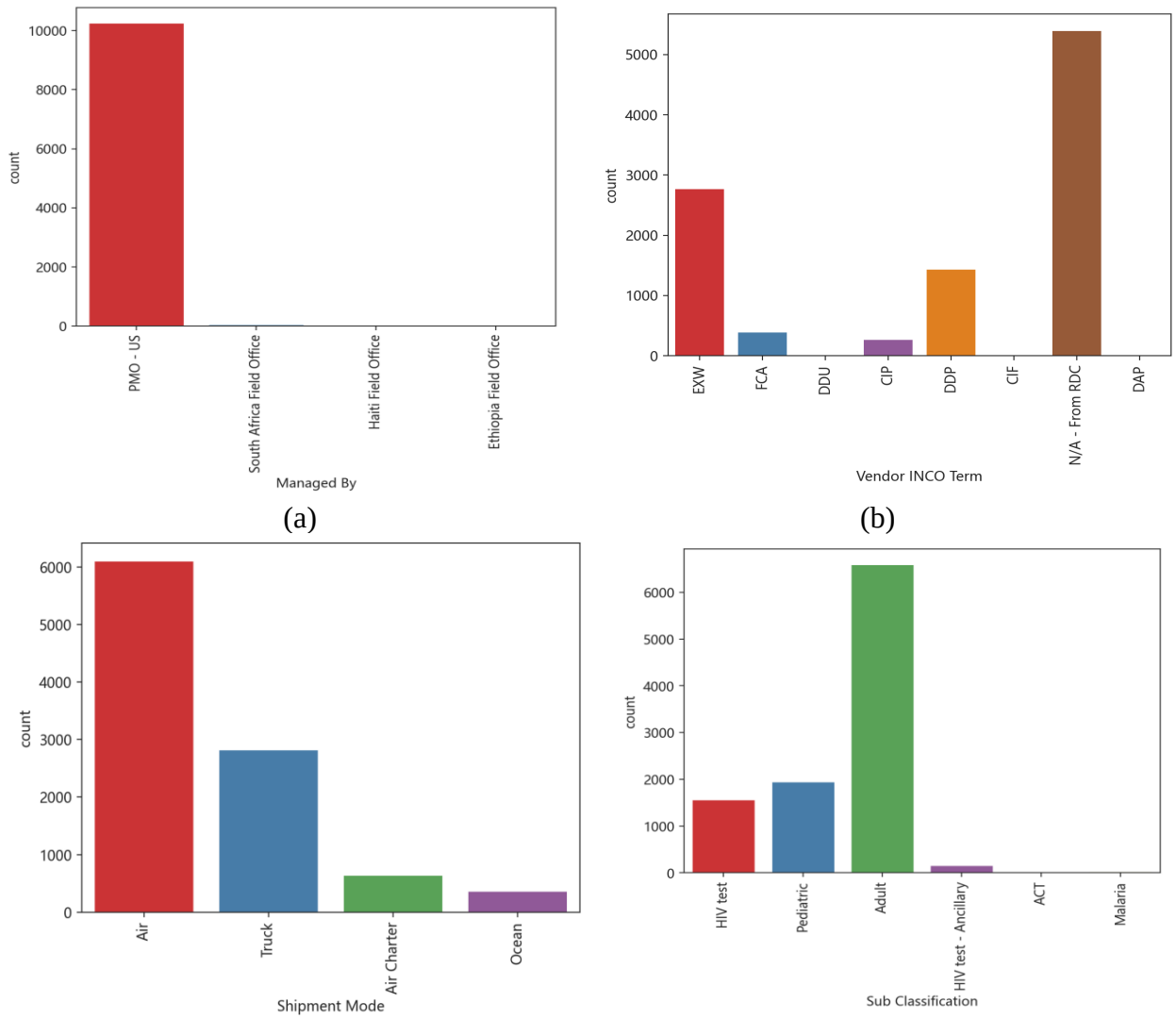

Figure 3. Sample size plots for each category in the partially discrete variable.

Observing the above Figure 2 (a)-(d), it is found that the value of managedby attribute column is more single and less useful for the prediction of logistics cost, so we choose to delete this attribute column, use the plural to fill the empty value of dosage attribute column, and at the same time delete all the rows with missing values in Shipment Mode column and reset the indexes to ensure the continuity of the indexes.

2.3. Data exploration

2.3.1. Logistics cost data analysis

In this paper, Freight Cost (USD) is used as the prediction target, and Figure 3 below shows the probability density plot of the distribution of the logistics cost data, and it is observed that the logistics cost data shows a skewed distribution, so the following considerations are made to deskew the prediction target.

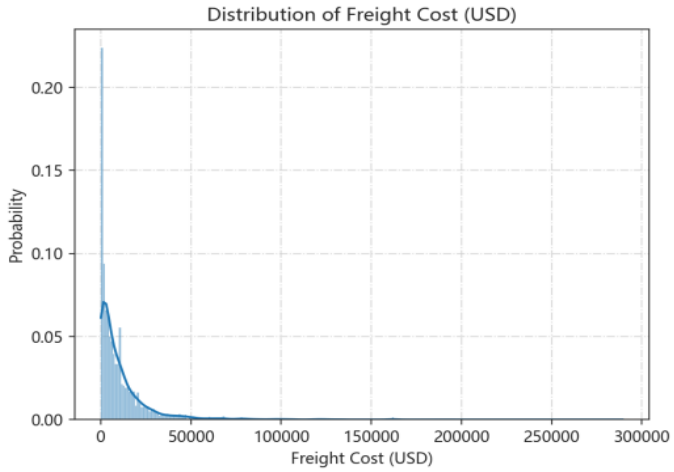

Figure 3. Probability density plot of logistics cost data distribution.

2.3.2. Exploration of factors influencing logistics costs

(1) Numeric variables

Exploring the distribution of numerical variables can help experiments to select appropriate characteristic variables as dependent variables to guide modeling. The results show that the skewed distribution of the data is so severe that the mean value may no longer be a good measure of central tendency. If the mean value is still used as a judgment criterion, it may lead to misjudgment of the location of the data centers, which in turn may mislead the decision making and increase the risk of error in the data analysis process. It is necessary to carry out the deskewing process to make the data closer to the normal distribution, so as to improve the accuracy and reliability of the analysis, and this study adopts the form of logarithmic transformation to deskew the data. The results are shown in Figure 4 below, the distribution of numerical variables after deskewing tends to be normal.

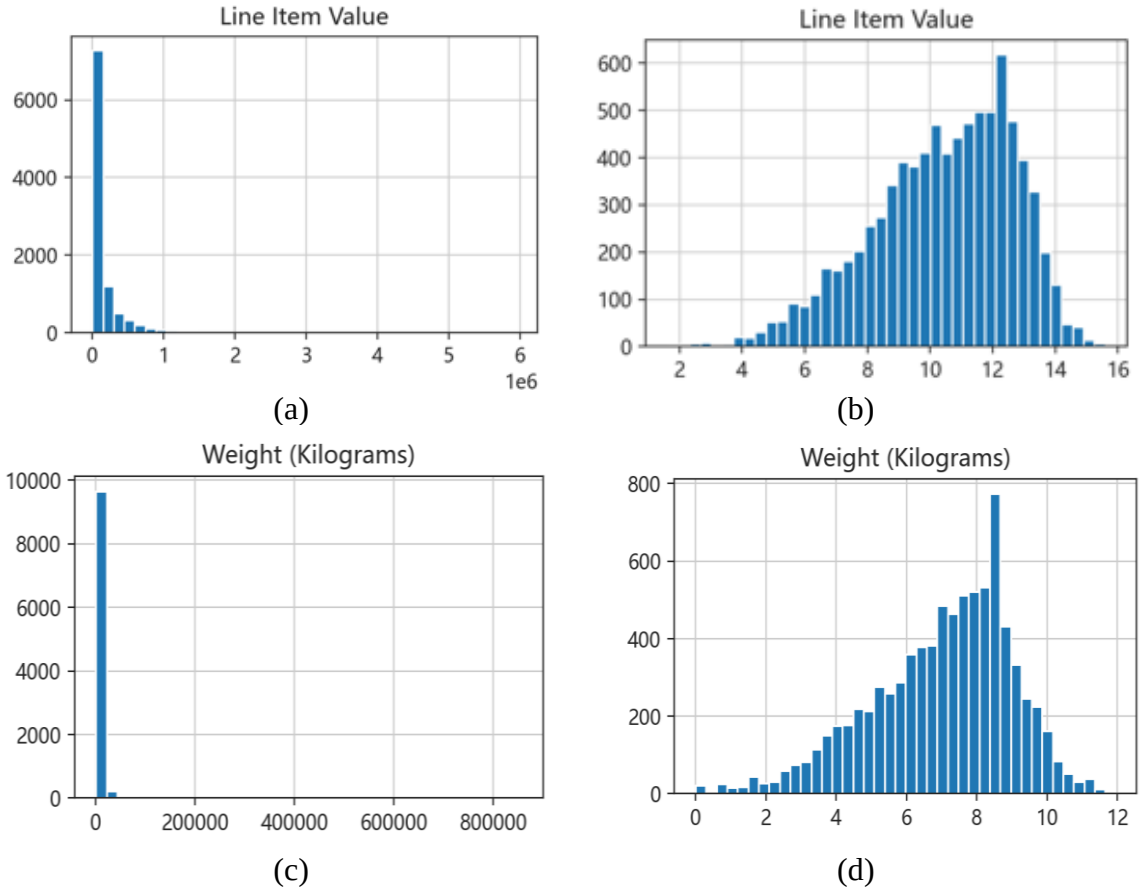

Figure 4. Probability plot of the distribution of some numerical variables.

(2) Discrete variables

All the discrete category features are identified, and the possible feature variables are displayed by means of box plots to explore the distribution state of the data, and the results are shown in the following figure, which is only partially displayed due to space reasons. From the box plot, it is found that there are more freight outliers and the data distribution is more concentrated. It is necessary to eliminate the outliers, using the interquartile method, by defining the range of outliers based on IQR to identify and eliminate the outliers in the data set. The results of the culled data are shown in Figure 5 (a)-(d) below. The outliers in the box plot are reduced after the quartile method treatment.

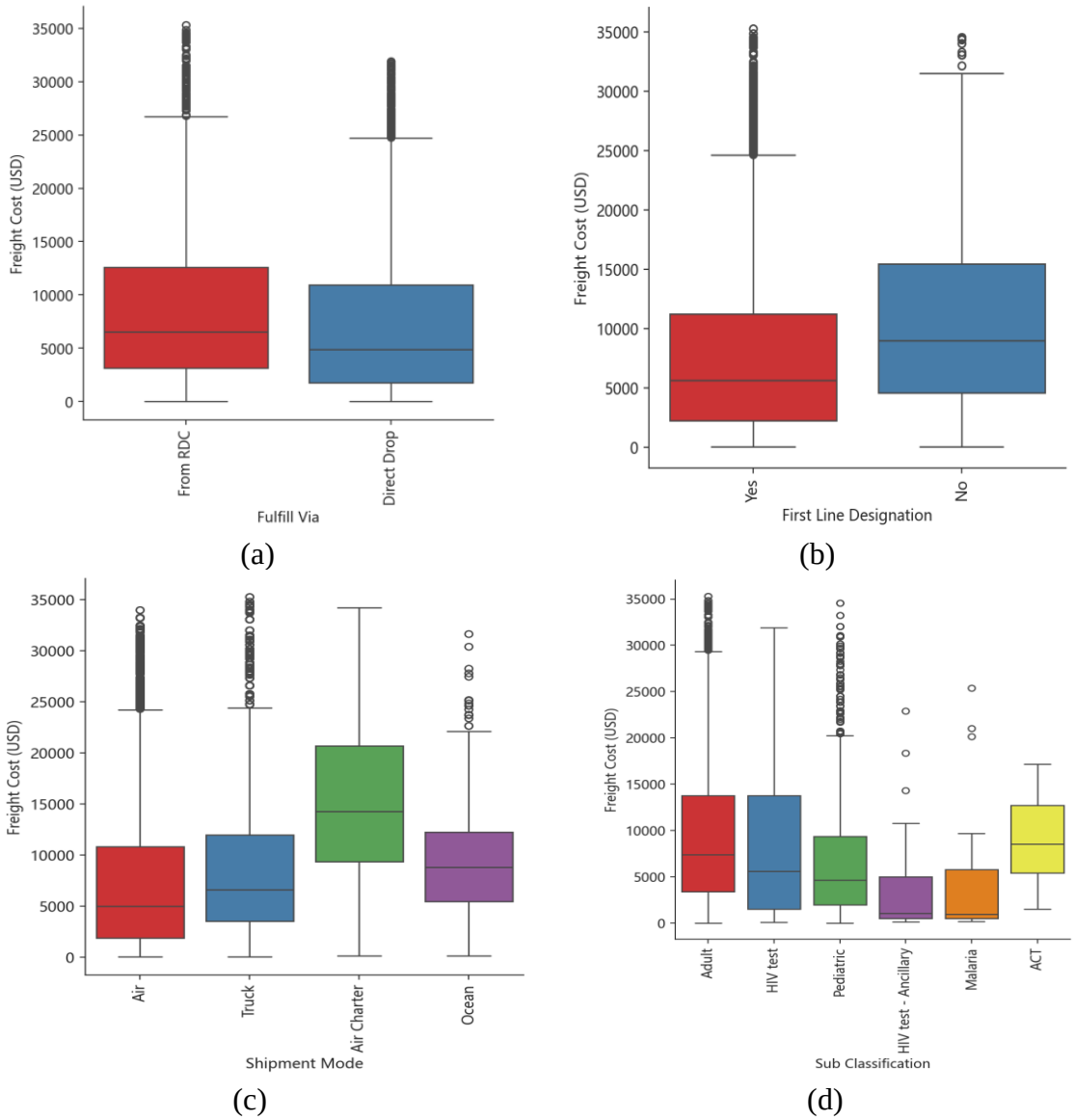

Figure 5. Box plot of data distribution for selected discrete variables.

(3) Correlation analysis

By quantifying the degree of association between two or more variables. By calculating the correlation coefficient, it is possible to understand whether there is a linear or non-linear relationship between the variables and the strength and direction of this relationship (positive or negative correlation) as shown in Figure 6.

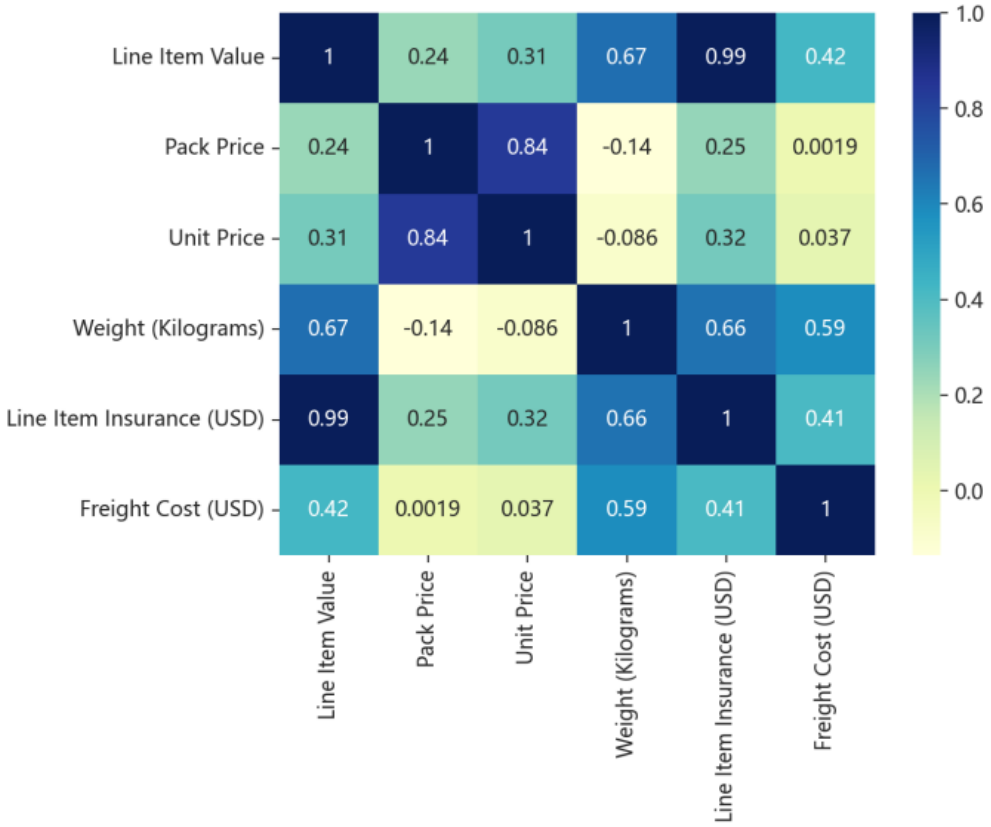

Figure 6. Correlation analysis of characteristic variables.

The correlation heat map shows a weak linear correlation between the above numerical type variables and logistics costs. Therefore, when building the forecasting model below, consideration should be given to trying to use a non-linear regression forecasting model.

3. Modeling

3.1. Coding discrete variables

The discrete variables in the already processed data were coded and converted into numerical data that the model could understand. Some of the coded data are shown in Table 2.

Table 2. Partially coded data.

PO / SO #	ASN/DN #	Country	...	Dosage	Dosage Form	Line Item Quantity	Line Item Value	Pack Price
5067	4275	15	...	31	10	564	8.4330	2.0980
5069	3645	39	...	4	11	8694	11.0718	2.0014
5083	4057	34	...	31	15	302	10.0924	4.3820
5093	4135	9	...	11	10	3145	9.1654	1.1118
2867	4471	23	...	19	7	338	8.8699	3.0469

3.2. Feature Screening

(i) Screening for discrete variables using chi-square tests.

In this study, the chi-square test [6] was used to select the four most important features in the discrete variables of the dataset, which have the strongest correlation with the target variables and may play an important role in the subsequent model training.

Finally Fulfill Via, Vendor INCO Term, Shipment Mode, Sub Classification features are selected by chi-square test.

(ii) Screening continuous variables using regression correlation.

The Pearson's correlation coefficient [7] is used as a scoring function to select the three numeric features (numeric_columns) that are most relevant to Freight Cost (USD). The names of these selected features were then extracted from the original feature set.

Three numerical features, Line Item Value, Weight (Kilograms), Line Item Insurance (USD), were finally selected.

3.3. Data set partitioning

In this paper, 80% of the data is divided into training set and 20% of the data is divided into test set for model evaluation using randomized division.

3.4. Lasso regression

Lasso regression model [8], is a linear regression method that restricts the model coefficients by adding L1 regularization terms. Its advantage lies in its ability to automatically perform feature selection and compress unimportant feature coefficients to zero, thus simplifying the model and improving interpretability while preventing overfitting. However, Lasso regression also has some drawbacks, such as the selection of regularization coefficients is more difficult and its computational complexity is higher compared to ordinary linear regression.

3.5. Random Forest Regression

Random forest regression modeling is an integrated learning method [9] that predicts continuous values by constructing multiple decision trees and outputting their average. It combines the prediction results of multiple weak learners to improve the accuracy and stability of the overall model. Advantages include the ability to handle high-dimensional data, automatic feature selection, and effective prevention of overfitting. However, the disadvantages of Random Forest are that the model may be overly complex, resulting in high computational costs, and it may be difficult to interpret the contribution of each feature to the prediction results in some cases.

3.6. Hyperparameter optimization

In this paper, we use grid search with cross-validation for model hyperparameter optimization. 5-fold cross-validation, and Table 3 below shows the optimal hyperparameters for lasso and random forest regression.

Table 3. Optimal hyperparameters for the four models.

Models (algorithms)	hyperparameterization
Unbiased lasso regression model	lasso__alpha=0.0
Unbiased lasso regression model	transformedtargetregressor__regressor__alpha=0.0
Unbiased random forest regression models	randomforestregressor__min_samples_leaf=3
Unbiased random forest regression model	transformedtargetregressor__regressor__min_samples_leaf=3

3.7. Comparison of model results

Table 4. Comparison of prediction errors of four models.

Models (algorithms)	MSE	MAE	R^2
Unbiased lasso regression model	34331361.66	4494.80	0.39
Unbiased lasso regression model	32572443.94	3988.94	0.41
Unbiased random forest regression models	10549172.20	1639.28	0.86
Unbiased random forest regression model	10590412.71	1584.59	0.86

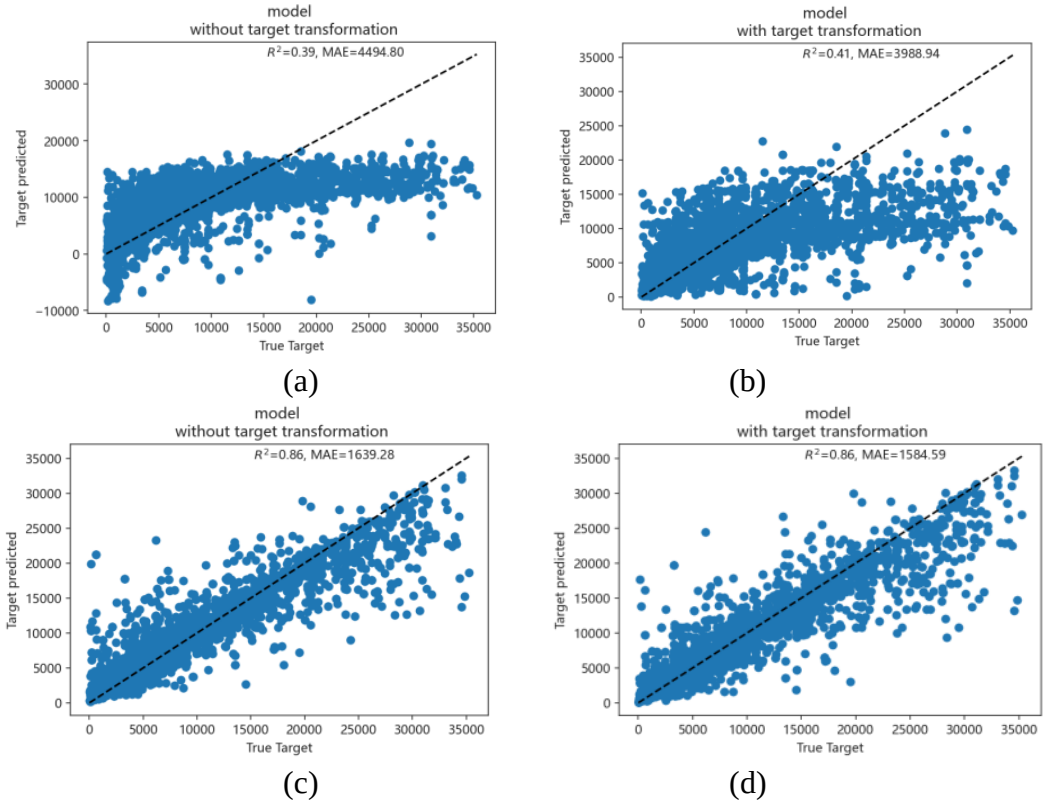

Figure 7. Scatter plot between true and predicted logistics costs.

Observing the above Table 4 and Figure 7 (a)-(d), it is found that the real value and the predicted value of lasso regression model are not linearly distributed, and there is a strong linear relationship between the predicted value and the true value in the random forest regression model. Compared with the lasso regression model, the MAE value of the random forest regression model decreases by 2855.52, and the R^2 increases by 0.47. The prediction error of the model decreases, and the prediction rate of the model increases. After the random forest regression model debias, R^2 has no significant change, but MAE value decreases by 54.69, and the model error decreases.

3.8. Model Interpretation

Model interpretation aims to elaborate how the model makes predictions based on the input data with respect to decision logic, feature importance, and model limitations [10]. It enhances model transparency, improves trust, accurately assesses prediction results, and identifies potential problems to optimize the model. The Random Forest Regression algorithm determines the importance of features by evaluating their contribution in reducing prediction error, and this study uses this library of algorithms to obtain feature scores, ranked to guide selection and optimization.

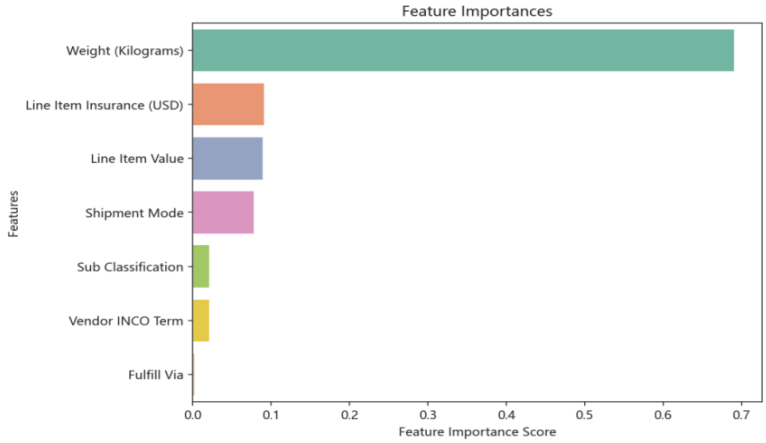

Figure 8. Ranking of model importance.

Observing the above figure 8, it is found that Weight(kilograms) has the highest feature importance score followed by features like Line term insurance(USD),Lin Item Value.

4. Conclusions

This paper explores the variables in the open source data by deskewing the skewed distributions in the numerical variables and treating the outliers in the discrete variables by quartile method. The chi-square test and Pearson's correlation coefficient were also used for feature selection of numerical and discrete variables, respectively. The lasso regression and random forest regression models were established respectively, and grid search and cross-validation were used to find the optimal hyperparameters of the models, optimize the models, and predict the data, and the results showed that the random forest regression model was better than the lasso regression model in terms of the results, and at the same time, the random forest regression model with deflated bias was better than that of the random forest model with no deflated bias, and finally, the importance of the characteristics of the logistic cost prediction was ranked. Interpretation of the model.

The core idea of fusion modeling, as an integrated learning technique, is to combine the prediction results of multiple base learners by some strategy with a view to obtaining superior performance over a single model. In the complex and variable task of logistics cost prediction, the fusion model can show significant advantages. Therefore, subsequent attempts can be made to optimize the prediction results by adopting the corresponding fusion model for prediction.

References

[1] Wang Lijuan, Li Xinyu. Coal logistics cost prediction based on PCA-EBP neural network [J]. World Science and Technology Research and Development, 2016, 38(05):1101-1106.

[2] Sun Changfeng. Research on logistics cost prediction based on gray correlation theory [J]. Logistics Technology, 2014, 33(23):242-244.

[3] Yanqiao Feng. Ship logistics and distribution cost prediction based on neural network [J]. Ship Science and Technology, 2020, 42(08):193-195.

[4] Huang Yuyi, Ai Xiaoqing, Wu Panyu. Social logistics cost prediction based on mixed-frequency data [J]. Statistics and Decision Making, 2020, 36(13):179-183.

[5] Lv Murong, Jia Lili, Hao Jingjing, et al. Research on countermeasures to reduce logistics cost in Yunnan Province based on integrated transportation big data [J]. Logistics Research, 2024, (04):42-53.

[6] Wei Zhidong, He Shumin, Huang Yuxin. Quantitative analysis of contract performance risk of enterprises during the epidemic--Statistical analysis based on the chi-square test of SPSS two-way unordered RxC table data of 392 cases in Guangdong Province [J]. Western Journal, 2021, (04):51-53.

[7] She Zihang, Xu Jiahua, Yao Zhiyu, et al. Analysis of online shopping big data based on Pearson's correlation coefficient--Taking Tmall Bai Runju flagship store transaction records as an example [J]. Journal of Hanshan Normal College, 2020, 41(03):16-22.

[8] Ye Wuyi, Xu Yincong, Jiao Shoukun. Statistical inference for multinomial adaptive Lasso regression [J]. Applied Probability Statistics, 2024, 40(01):107-121.

[9] Zheng Haitao. Prediction of coal railroad shipment based on adaptive boosting random forest algorithm [J]. China New Technology and New Products, 2023, (01):70-72.

[10] Yang Bo, Wang Zhengbing. Study on the Construction of Agricultural Products Logistics Network in Gansu Province under the Background of Rural Revitalization Strategy--Based on the Explanation of Logistics Gravity Model [J]. Journal of Central South Forestry University of Science and Technology (Social Science Edition), 2021, 15(06):92-100.