ISSN 2616-5775 Vol. 7, Issue 2: 109-114, DOI: 10.25236/AJCIS.2024.070215

# Logistics Cost Prediction Based on Random Forest Model

**Feifei Xie[1], Wei Yu[2]**

[1]*School of Business, Anhui University, Hefei, 230601, China*
[2]*School of Automotive and Transportation Engineering, Jiangsu University, Zhenjiang, 212013, China*

***Abstract:*** *Predicting logistics costs is imperative for enterprises to effectively manage and make informed decisions regarding logistics expenditures. This paper employs a systematic approach to this end. Initially, a Pearson correlation analysis is conducted on variables including Line Item Quantity, Line Item Value, Weight, and Freight Cost, revealing a strong correlation between Weight and Freight Cost. Subsequently, a chi-square test is applied to variables such as Fulfill Via, Vendor INCO Term, Shipment Mode, Product Group, Sub Classification, and Freight Cost. This analysis identifies Vendor INCO Term, Shipment Mode, and Sub Classification as pivotal influencing factors. To further enhance predictive accuracy, K Nearest Neighbour Regression (KNN), Support Vector Regression (SVR), and Random Forest (RF) models are individually employed for logistics cost prediction. Comparative analysis of prediction errors indicates that the RF model outperforms KNN, SVR, and other models, showcasing superior logistics cost prediction accuracy.*

***Keywords:*** *Logistics Costs, Random Forest, Chi-square Test, Support Vactor Regression, K-Nearest Neighbor*

## 1. Introduction

Modern logistics is the "connector" of economic development, one end is connected to the production, one end is connected to the consumption, to promote the realisation of the market supply and demand docking, accelerate the physical circulation of goods, and promote economic development. However, for most enterprises, logistics costs account for a large proportion of total costs, the control and prediction of logistics costs has become an important task of logistics management. Research on the prediction of logistics costs is conducive to providing a scientific basis for logistics cost decision-making, improving logistics and transport efficiency, and reducing logistics and transport costs.

Sun Shusheng et al. (2014) analysed the national macro logistics cost data by establishing a reasonable multiple linear regression model, and the results show that the equation is in line with the reality, and can be used to predict the macro logistics cost[1].Chen Longtao et al. (2015) set up a principal component regression prediction model for the forecasting of coal logistics cost, and predicted the logistics cost of a coal producer in Ordos, and the results show that the model predicts the results well[2].Yang Jing et al. (2017) predicted coal logistics cost based on improved support vector regression machine, and the results showed that the model had high prediction accuracy[3]. Tong Linlin (2023) used BP neural network model to predict cross-border e-commerce logistics cost, and the results found that the model had high accuracy, and it could effectively improve the efficiency of e-commerce logistics prediction[4].

In the paper, various machine learning algorithms, including support vector regression, random forest, and K-nearest neighbor regression, were employed for logistics cost prediction. Through a comparative analysis of prediction outcomes, the random forest algorithm emerged as the optimal model, demonstrating higher prediction accuracy than other models. This finding attests to the efficacy of the random forest model in predicting logistics costs. Notably, in contrast to neural networks and other "black box" machine learning algorithms, the random forest model offers the advantage of reducing feature impurity in tree splitting. This reduction facilitates the determination of feature importance, providing a structured representation of the variables' contribution to the prediction outcome. The ordered importance of variables allows for the interpretability of the model, distinguishing it from less transparent "black box" algorithms.

ISSN 2616-5775 Vol. 7, Issue 2: 109-114, DOI: 10.25236/AJCIS.2024.070215

## 2. The basic funamental of Random Forest Algorithm

Random forest algorithm is a data mining and machine learning model with decision tree as the base learner. The algorithm requires fewer parameters to be tuned and has an advantage in solving large-scale samples or problems with multiple categorical data. Random forest algorithm processing flow is as follows [5-8]:

(1) Multiple samples are drawn from the original samples using Bootstrap method with put-back to generate multiple sampling sets, each of which forms a decision tree consisting of two kinds of data that are sampled and not sampled (out-of-bag data);

(2) Splitting the decision tree by selecting the optimal features based on the Gini metrics to maximise the growth of each decision tree;

(3) Re-modelling based on the prediction error of the out-of-bag data to determine the optimal number of decision trees;

(4) According to the prediction value given by each decision tree, the prediction results of classification and regression problems are given using voting or averaging methods respectively.

## 3. Example Analysis

### *3.1 Data Source and Classification*

The data in the paper comes from https://www.kaggle.com/datasets/lakhankumawat/real-data-on-logistics, and the data related to logistics cost prediction include Fulfill Via, Vendor INCO Term, Shipment Mode, Product Group, Sub Classification, Line Item Qua ntity, Line Item Value, Weight, Freight Cost and other nine variables. Among them, Line Item Quantity, Line Item Value, Weight, and Freight Cost are numerical variables, and Fulfill Via, Vendor INCO Term, Shipment Mode, Product Group, and Sub Classification are discrete variables. The following are discrete variables: Line Item Quantity, Line Item Quantity.

The histograms of four of the numerical variables such as Line Item Quantity, Line Item Value, Weight, and Freight Cost are shown in Figure 1 below, and the results show that there are no significant outliers for the variables.

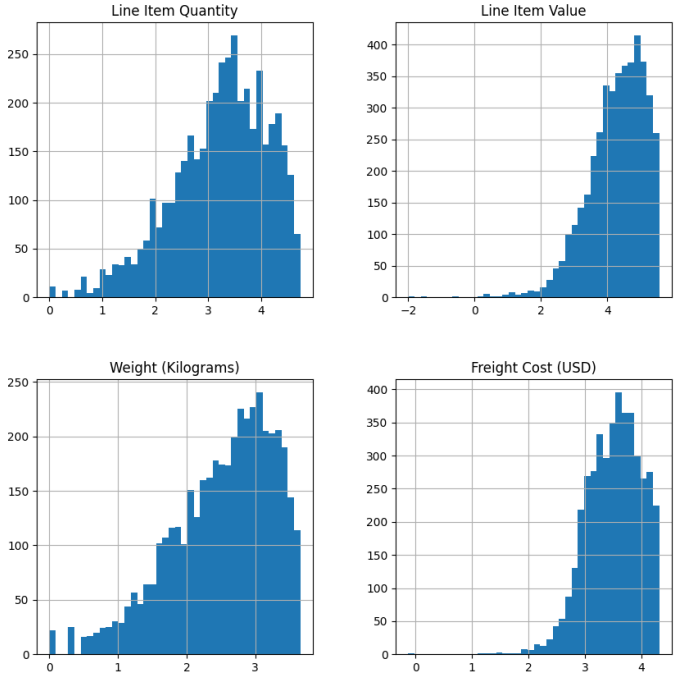


*Figure 1: Histogram of numerical variables*

The discrete variables Fulfill Via include Direct Drop, From RDC, Vendor INCO Term includes EXW, FCA, CIP, DDP, CIF, N/A-From RDC, DDU, DAP, Shipment Mode includes Air, Truck, Air Charter,

ISSN 2616-5775 Vol. 7, Issue 2: 109-114, DOI: 10.25236/AJCIS.2024.070215

Ocean, Product Group includes HRDT, ARV, ACT, MRDT, ANTM, Sub Classification includes HIV test, Pediatric, Adult, HIV test, and HIV test. In Shipment Mode, there are Air, Truck, Air Charter, Ocean, Product Group includes HRDT, ARV, ACT, MRDT, ANTM, Sub Classification includes HIV test, Pediatric, Adult, HIV test-Ancillary, ACT, Malaria.

### *3.2 Data analysis*

Data exploration was conducted for all data variables before prediction, including Pearson correlation analysis for numerical variables and box plot analysis for discrete variables.

#### *3.2.1 Pearson correlation analysis of numerical variables*

Analysing Figure 2 shows that the correlation coefficients of Line Item Quantity, Line Item Value, Weight variable and Freight Cost variable are all greater than 0.5, indicating that each of them is positively and moderately correlated with each other. Among them, the correlation coefficient between Weight variable and Freight Cost variable is greater than 0.6, which is relatively strong[9].

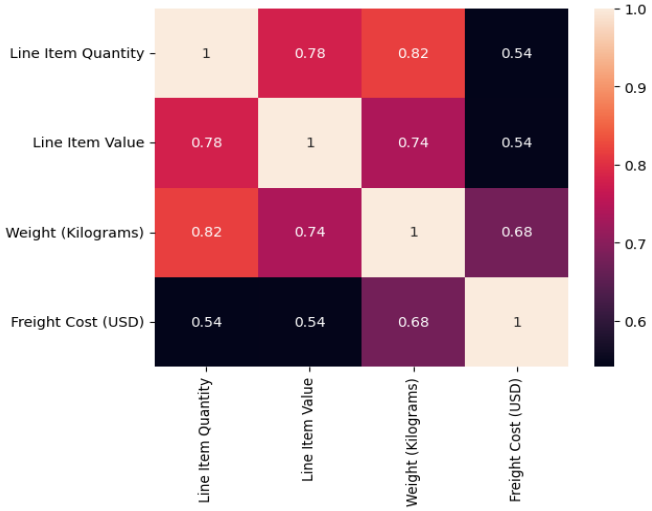


*Figure 2: Pearson correlation analysis results chart*

#### *3.2.2 Box plot analysis of discrete variables*

The examination of Figures 3 through 7 reveals two discrete categories under the variable "Fulfill Via." These categories exhibit a similar degree of concentration, concentration trend, and skewness for various values. Notably, "From RDC" displays smaller variance. Conversely, the discrete variable "Vendor INCO Term" encompasses eight categories, exhibiting substantial disparities in concentration, concentration trend, skewness, and variance across different values. This observation suggests evident distinctions in logistics costs among categories, including "EXW," "FCA," and "CIP," which demonstrate a higher degree of concentration. The discrete variable "Shipment Mode" comprises four categories, showcasing considerable variation in concentration and variance among different values, characterized by a high and normally distributed concentration trend. Moreover, the discrete variable "Product Group" comprises five categories, displaying significant variability in concentration and variance among different values. Notably, the variable "ACT" exhibits the smallest variance, while the variable "MRDT" demonstrates a left-skewed distribution. Similarly, the discrete variable "Sub Classification" includes six categories, featuring considerable variation in variance and concentration trend among different values. Remarkably, "ACT" exhibits the smallest variance, while "Malaria" displays a right-skewed distribution.

The chi-square test serves to assess the independence between labels and discrete features. When independence is established, it implies that the features do not significantly contribute to label prediction. In this study, the CHI2 method is employed to filter discrete variables influencing logistics costs. The selected variables, namely Vendor INCO Term, Shipment Mode, and Sub Classification, emerge as more influential factors, signifying a substantial impact on logistics costs. This aligns with the findings of the box plot analysis, reinforcing the importance of these three variables in predicting and understanding logistics costs.

ISSN 2616-5775 Vol. 7, Issue 2: 109-114, DOI: 10.25236/AJCIS.2024.070215

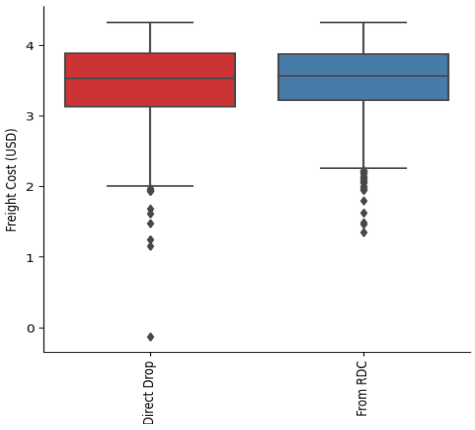


*Figure 3: Variable Fulfill Via box diagram*

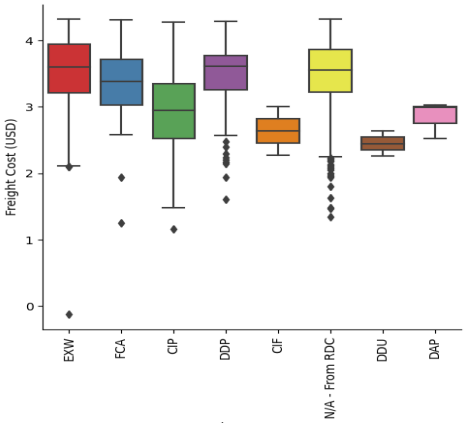


*Figure 4: Variable Vendor INCO Term box diagram*

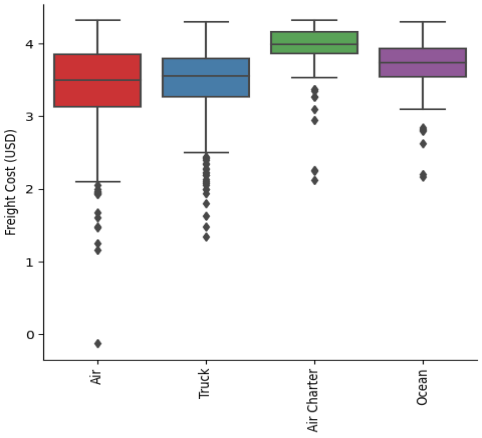


*Figure 5: Variable Shipment Mode box diagram*

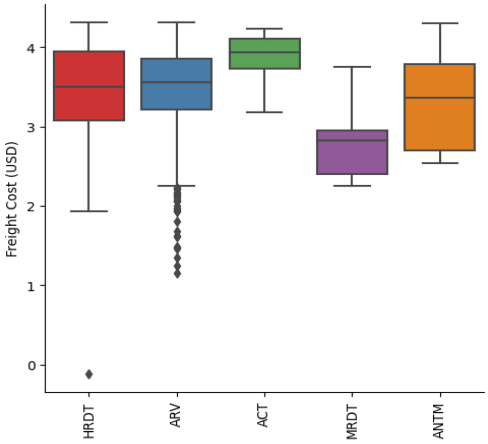


*Figure 6: Variable Product Group box diagram*

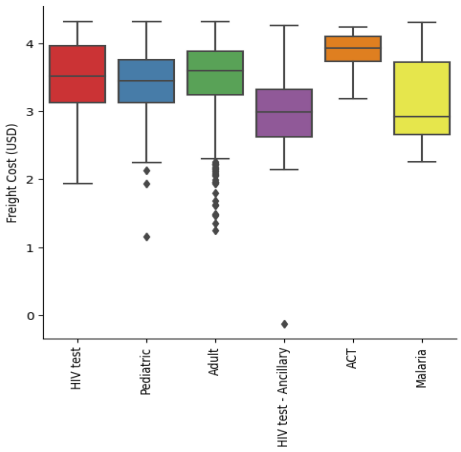


*Figure 7: Variable Sub Classification box diagram*

### *3.3 Establishment of Random Forest Prediction Logistics Cost Model*

80% of all data are randomly selected as the training set to train the model, and 20% of the data are used as the test set. In order to prove the advantage of machine learning RF prediction model, this paper will use KNN, SVR, RF three prediction models to conduct experiments and compare the experimental results.

KNN algorithm parameter setting: n_neighbors=17, SVR algorithm parameter setting: gamma=0.01,C=100, RF algorithm parameter setting: max_depth=10,n_estimators=70, and simulation is carried out by using python, in which the prediction results of Random Forest are shown in Figure 8.

ISSN 2616-5775 Vol. 7, Issue 2: 109-114, DOI: 10.25236/AJCIS.2024.070215

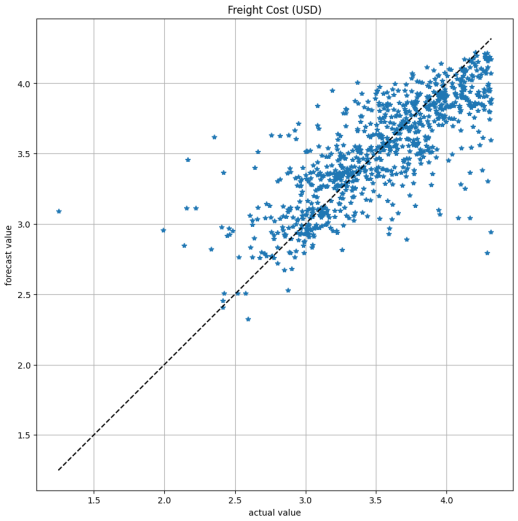


*Figure 8: RF model prediction result*

The results of three different prediction models are compared and the results are shown in Table 1.

*Table 1: Comparison of prediction errors of several models*

| Model | MSE | MAE |
|---|---|---|
| KNN | 0.089 | 0.212 |
| SVR | 0.096 | 0.223 |
| RF | 0.086 | 0.207 |

### *3.4 Analysis of results*

Analysing Table 1 shows that:

(1) The RF model has the smallest MSE and MAE values, indicating that the model has the smallest prediction error and the highest prediction accuracy, which is better than the KNN model and the SVR model.

(2) SVR model is more dependent on data, weak in generalisation and has the worst prediction accuracy.

### *3.5 Random forest feature importance ranking*

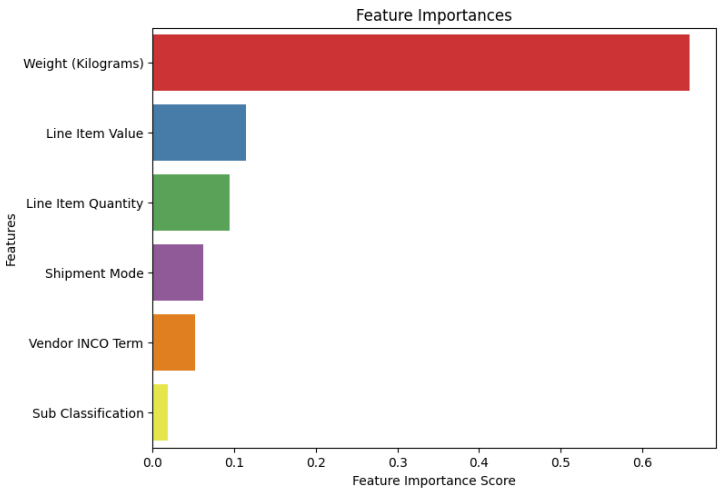


*Figure 9: Ranking of feature importance*

The Random Forest Model makes interpretability based on the contribution of variables to the prediction results and can rank the degree of importance of variables. In this paper, the random forest

ISSN 2616-5775 Vol. 7, Issue 2: 109-114, DOI: 10.25236/AJCIS.2024.070215

model with the highest prediction accuracy was used to predict logistics costs, and all the factors affecting the prediction of logistics costs were ranked, and the results showed that the Weight variable had the greatest influence on logistics costs, followed by Line Item Value, Line Item Quantity, Shipment Mode, Vendor INCO Term, Sub Classification, as shown in Figure 9. Therefore, if enterprises want to control logistics costs, they need to focus on Weight, Line Item Value and Line Item Quantity.

## 4. Conclusion

In this paper, for the problem of logistics cost control and prediction, using a variety of self-learning algorithms such as KNN, SVR, RF, combined with a variety of methods such as chi-square test, Pearson correlation test and other methods, prediction simulation was carried out using existing historical data. The results show that RF logistics cost prediction model is significantly better than KNN, SVR and other models in terms of prediction accuracy. Therefore, RF prediction model can effectively predict logistics costs.

However, the RF model has a certain degree of randomness and fewer adjustable parameters, and when there is too much data, integrated learning algorithms such as XGboost, which has more hyperparameters and faster iteration speed, can be used. In addition, the number of influencing factors of logistics cost discussed in this paper is relatively small, and future research can further expand the scope of influencing factors to further improve the accuracy of model prediction.

## References

*[1] SUN Shusheng, LUO Baohua. Application of multiple linear regression model in logistics cost prediction [J]. Business Age, 2014(18):19-21.*
*[2] CHEN Longtao, LU Shichang,TAI Xiaohong et al. Research on coal logistics cost prediction based on principal component regression analysis [J]. Resource Development and Markets,2015,31(06):641-644.*
*[3] YANG Jing,LI Junfu, ZHANG Gaoqing. Coal logistics cost prediction based on improved support vector regression machine [J]. Journal of Guangxi University (Natural Science Edition), 2017, 42(04): 1623-1627.*
*[4] Tong LL. Cross-border e-commerce logistics cost prediction based on neural network [J]. Logistics Technology, 2023, 46(04):48-51.*
*[5] Su Yuteng, Lv Siyun, Xie Wenhan et al. Analysis of risk factors for the development of type 2 diabetes mellitus based on LASSO regression and random forest algorithm [J]. Journal of Environmental Hygiene, 2023, 13(07): 485-495.*
*[6] LIU Fuqiang, CHEN Xiaodong, LI Shengfu et al. Prediction of permeability coefficient of sand body based on random forest regression [J]. Uranium Ore Geology, 2023, 39(04): 653-661.*
*[7] Ding Sha; Shen Taorong; Zhang Yanfei; Du Huanzhe; Wu Yu; Zou Xiaoyong. A study on category identification model of tobacco extracts based on random forest algorithm [J]. Journal of Analytical Testing, 2023, 42(11):1510-1516.*
*[8] Zou, Jie; Li, Lu. A study on stock price prediction based on SA-BiGRU model with random forests [J]. Commodity Prices in China, 2023, (11):52-56.*
*[9] GUO Liang, GUO Zixue, JIA Hongtao et al. Identification of residential electricity theft based on Pearson correlation coefficient and SVM [J]. Journal of Hebei University (Natural Science Edition), 2023, 43(04):357-363.*