DOI: 10.53469/wjimt.2023.06(05).07

Research on Urban E-commerce Logistics Volume Based on XGBoost Regression Tree

Shiyi Huang[1, #], Minghao Wang[2, #], Tingting Chen[2, #]

[1]Industry College of Blockchain, Chengdu University of Information Technology, Chengdu, Sichuan, China
[2]College of Applied Mathematics, Chengdu University of Information Technology, Chengdu, Sichuan, China
[#] These authors contributed equally to this work.
[1]huangshiyivvv@outlook.com, [2]1992879928@qq.com, [3]yilantingting@qq.com

Abstract: ***With the popularization of the Internet, online shopping has become prevalent, leading to rapid development in the logistics industry. In order to predict the volume of logistics between cities, this paper proposes a city logistics volume prediction model based on the XGBoost algorithm and validates it using real city logistics datasets. The experiment demonstrates that this prediction model can accurately forecast the logistics volume between different cities, exhibiting good predictive performance and robustness. Research on logistics volume prediction is beneficial for optimizing urban logistics planning and enhancing logistics efficiency.***

Keywords: Urban logistics; Logistics volume prediction; XGBoost.

1. INTRODUCTION

The rapid expansion of e-commerce in China has propelled logistics into a pivotal position within modern society.

Yang et al. applied the RBF neural network to predict the volume of express delivery services [1]. Xu et al. utilized an improved PSO-BP algorithm for express delivery volume prediction [2].

The optimization of parameter tuning for previous studies has posed significant challenges, characterized by the high demand for quality of data and the inherent risk of converging towards local optima. Therefore, this paper proposes an efficient and accurate prediction model capable of handling large, complex, and nonlinear data in urban logistics.

The aim of this study is to predict urban logistics volume by applying mathematical modeling and algorithm optimization techniques, thereby improving logistics planning and decision-making. This will enhance the operational efficiency of major express delivery companies, optimize resource allocation, and increase customer satisfaction.

2. DATA SELECTION AND PRE-PROCESSING

2.1 Data Selection

In this study, a domestic delivery company in China is selected as the research object. Data on express transportation between 25 cities (denoted as City A to City Y) from April 19, 2018, to April 17, 2019, were collected, including shipment dates, delivering cities, receiving cities and express delivery quantity. The city names have been replaced with letters. The objective is to predict the quantity of express deliveries between certain "delivering city to receiving city" city pairs on April 18th and April 19th, 2019. A portion of the data is presented in Table 1.

Table 1: A portion of the data

Date	Delivering city	Receiving city	Express delivery quantity (PCS)
2018/4/19	S	Q	42
2018/4/20	L	G	141
2018/10/18	O	R	0
2019/1/18	V	G	153
2019/4/13	L	K	286
2019/4/16	M	U	67

2.2 Data Pre-processing

Firstly, the substitution of alphabetic characters for city names poses certain challenges in data processing. To mitigate this adverse impact, we employ one-hot encoding to transform city names into numerical representations. Subsequently, the dataset is partitioned into training, validation, and test sets. The training set is used for hyperparametric tuning and model evaluation, while the test set is reserved for the final assessment of the model's performance.

3. MODEL ESTABLISHMENT

3.1 Model Principles

The XGBoost algorithm is a machine learning algorithm based on the gradient boosting framework. Unlike traditional gradient boosting decision trees, XGBoost introduces a regularization term to the loss function and utilizes the second-order Taylor expansion of the loss function for fitting [3]. XGBoost is characterized by its efficiency, flexibility and accuracy, making it widely applied in various domains such as data mining, prediction and regression.

For a given dataset with n instances and m features, the XGBoost model can be represented as:

$$\mathrm{y}_1 = \sum_{k=1}^{K} f_k\left(X_i\right), \quad f_k \in F \quad (i = 1, 2, \ldots, n) \tag{1}$$

In the formula, y_1 represents the predicted target value of the model, K represents the number of base classifiers, and $f_k(X_i)$ represents the prediction result of the k-th base classifier on the samples.

For this model, we seek the optimal parameters based on the principle of minimizing the objective function to establish the optimal model [4]. The objective function of XGBoost consists of two parts: the loss function L and the regularization term Ω.

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_J} g_i\right)^2}{\sum_{i\in I_J} h_i + \lambda} + \gamma T. \tag{2}$$

According to our experimental objective, we choose the squared error loss function, defined as follows:

$$\mathbf{L} = \sum_{i=1}^{n}\left(\mathrm{y}_i - \hat{y}_i\right)^2 \tag{3}$$

In the formula, y_i is the actual value and $\hat{y}_i$ represents the predicted target value of the model. Additionally, based on the complexity of the data, we incorporate a regularization term to prevent overfitting.

$$\Omega = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{4}$$

In this formula, γT represents the L1 regularization term and $\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2$ represents the L2 regularization term. At this point, we introduce a new function into the model that aims to minimize the objective function as much as possible. The specific process is as follows:

$$\begin{gathered}\hat{y}_l^{(0)} = 0. \\ \hat{y}_l^{(1)} = \hat{y}_l^{(0)} + f_1\left(x_i\right). \\ \hat{y}_l^{(2)} = \hat{y}_l^{(1)} + f_2\left(x_i\right)\cdot\end{gathered} \tag{5}$$

$$\dots\dots$$

$$\hat{y}_l^{(t)} = \hat{y}_l^{(t-1)} + f_t\left(x_i\right).$$

At this stage, the objective function is represented as:

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} \left(\text{y}_i - \left(y_i^{(t-1)} + f_t\left(x_i\right) \right) \right)^2 \tag{6}$$

After performing Taylor expansion and removing the constant term, with the known structure part q of the tree, we can utilize the objective function to search for the optimal w_j and obtain the optimal value of the objective function. In essence, this can be categorized as a problem of minimizing a quadratic function. The solution is derived as:

$$w_j^* = \frac{-\sum_{i \in I_J} g_i}{\sum_{i \in I_J} h_i + \lambda} \tag{7}$$

$$Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_J} g_i \right)^2}{\sum_{i \in I_J} h_i + \lambda} + \gamma T. \tag{8}$$

3.2 Model Training

By recursively invoking the aforementioned tree construction method, we generate a multitude of regression tree structures. We employ the Obj function to search for the optimal tree structure and integrate it into the existing model to establish the optimal regression model. Through parameter tuning, we obtain the loss curve of the model as depicted in Figure 1.

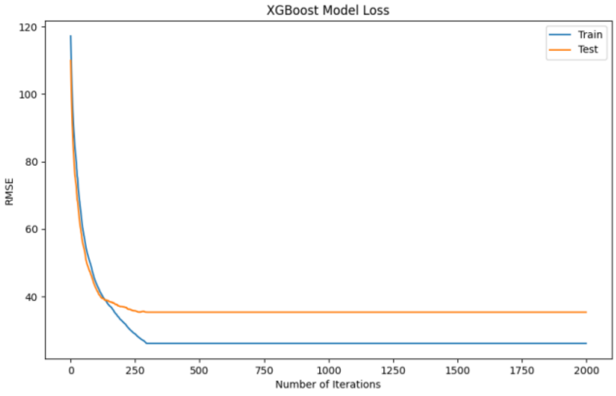


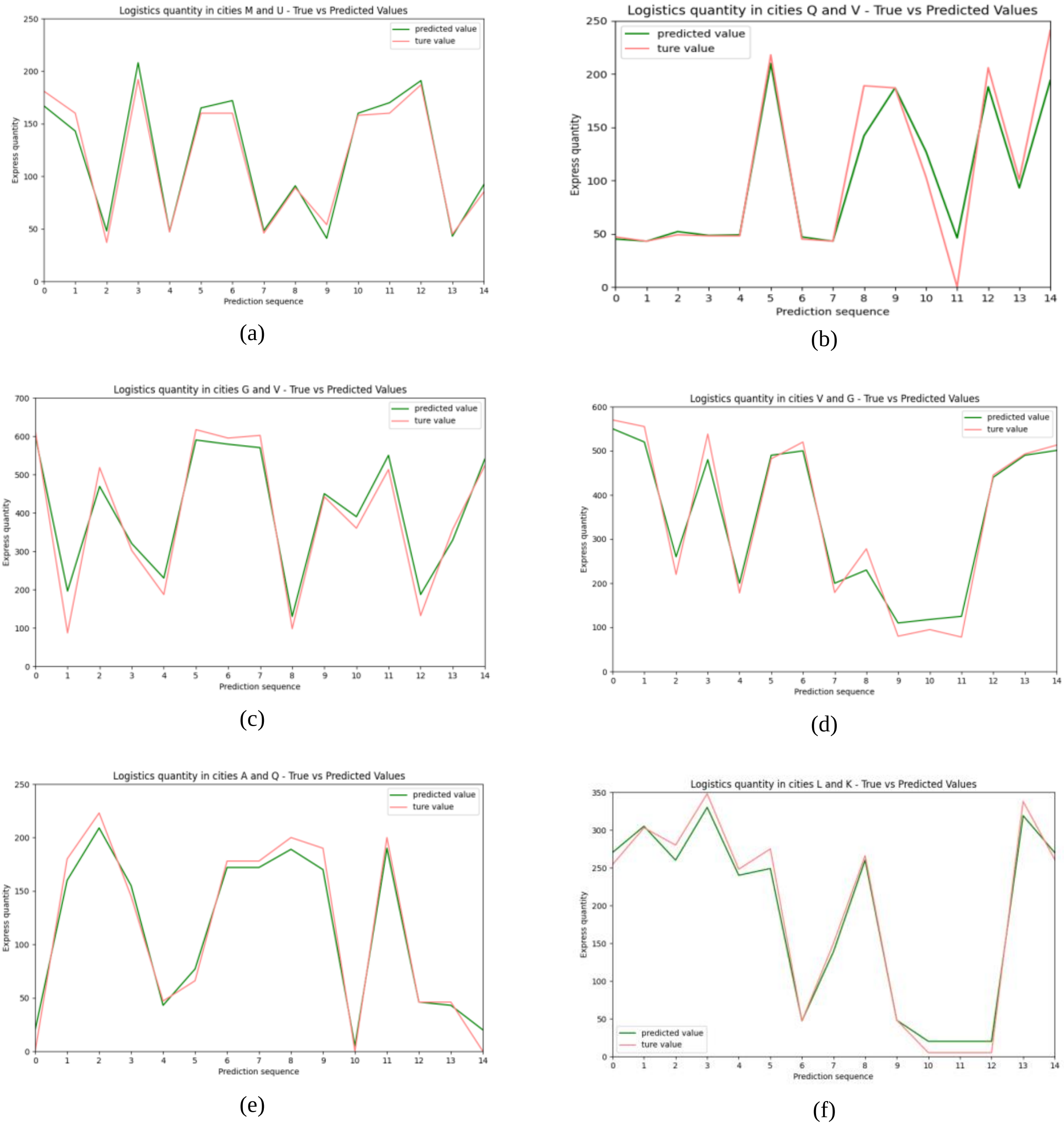
Figure 1: XGBoost model loss curve

From the obtained model loss curve, it can be observed that the model has shown good performance.

4. MODEL SOLUTION

4.1 Regression Fitting

Based on the established XGBoost regression model, the fitting effect on the test set is observed, and the fitting graph for the prediction of logistics volume between city pairs is shown in Figure 2:



(a) (b) (c) (d) (e) (f)

Figure 2: Regression fitting curve

Upon observation, it can be seen that the model fitting results are satisfactory, as it can effectively predict key features such as inflection points in the data.

4.2 Result of the Model

The predicted results for the logistics volume between certain site cities can be seen in Table 2.

Table 2: The predicted results of the model

Date	The logistics volume between the "delivering city to receiving city" city pairs	
2019/4/18	M-U	86.70854187011719
	Q-V	45.5
	G-V	672.9294171276058

2019/4/19	V-G	544.9572185015046
	A-Q	105.97313555325515
	L-K	112.5

5. MODEL EVALUATION

The XGBoost model is evaluated using various metrics to assess its performance. Metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R² are employed to evaluate the predictive accuracy of the model. The evaluation results are presented in Table 3.

Table 3: Evaluation of XGBoost model prediction

	MSE	RMSE	MAE	MAPE	R²
Training set	368.38	19.193	10.132	9.29	0.902
Cross-validation set	2109.676	42.516	29.463	25.63	0.404
Test set	702.114	26.497	21.171	24.538	0.82

Through the evaluation metrics of the cross-validation set, the hyperparameters can be continuously adjusted to obtain a reliable and stable model [5]. As the R² value approaches 1, the model accuracy increases. Based on these evaluation results, it can be inferred that the model fits well to this data.

6. SUMMARY

This paper proposes a city logistics volume prediction model based on the XGBoost algorithm and validates it using real data. The experimental results demonstrate that the model accurately predicts the logistics volume between cities and exhibits good predictive performance and robustness.

In the future, it is possible to further expand the scale of the dataset, optimize model parameters, improve prediction accuracy, and apply the model to practical urban logistics management, thereby promoting the efficient development of urban logistics.

REFERENCES

[1] Yang Yue: Research on Yunda Express Business VolumeForecast Based on RBF Neural Neywork (Master of Engineering, Harbin University of Science and Technology, China 2022).

[2] Xu Rongbin, Wang yeguo, Wang futian, et al. Prediction of package volume based on improved POS-BP [J]. Computer Integrated Manufacturing Systems, 2018, 24 (7): 9.

[3] Chen Tianqi, C. Guestrin. "XGBoost: A Scalable Tree Boosting System." ACM (2016).

[4] Friedman, Jerome H. Greedy Function Approximation: A Gradient Boosting Machine [J]. Annals of Statistics, 2001, 29 (5): 1189-1232.

[5] Ron, K. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995 American Association for Artificial Intelligence, 1995.

Author Profile

Shiyi Huang, undergraduate student of Chengdu University of Information Technology. Research direction: Blockchain.

Minghao Wang, undergraduate student of Chengdu University of Information Technology. Research direction: Information and Computing Science.

Tingting Chen, undergraduate student of Chengdu University of Information Technology. Research direction: Information and Computing Science.