Numerical Algebra and Scientific Computing (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/nasc.2022.010101
ISSN 2523-6431 Vol. 1 Num. 1

# *Analysis of Domestic Rebar Demand Based on Pearson Correlation Coefficient and XGBoost Model*

**Junshan Huang[1], Longfei Lu[1], Hongzhi Chen[1], Yan Liang[2,*]**

[1]*School of Computer, Guangdong University of Petrochemical Technology, Maoming, Guangdong, China*
[2]*School of Science, Guangdong University of Petrochemical Technology, Maoming, Guangdong, China*
[*]*Corresponding author*

***Keywords:*** Demand for rebar, PEARSON correlation, Gaussian distribution, XGBoost model

***Abstract:*** Rebar is one of the largest steel products in China. Rebar is widely used in building, bridge, road and other civil engineering construction. It is an indispensable structural material for infrastructure construction. To grasp the demand dynamics of rebar in the market reasonably and effectively has great significance in practice. The prediction of rebar demand is conducive to deepening the supply-side structural reform of the rebar industry, as well as improving the supply and demand situation, and alleviating the overcapacity situation of the rebar industry. The investment strategy of rebar futures can be adjusted according to the prediction results of rebar demand. Many factors affect the demand of the rebar market. To accurately predict the demand of rebar, this paper will first preprocess the sample data and standardize the deviation. Then XGBOOST model is adopted to integrate multiple base decision trees. Since the decision tree has the characteristics of nonlinear fitting, it can accurately predict the demand of rebar and provide an effective method to solve the long-term backward construction foundation recommendation in China.

## 1. Introduction

The scientific and technological revolutions in the world have stimulated the rapid development of the world economy, as well as the development of the rebar industry. Since rebar is widely used in civil engineering construction, it is the largest product of steel in China for recent decades. In recent years, the fierce market has not been eliminated, and has led to serious inflation or overcapacity in reinforcement industries around the world.

With the "Belt and Road", Beijing-Tianjin-Hebei coordinated development and the development of the Yangtze River Economic Belt, the demand for rebar in transportation and infrastructure is growing rapidly, which affecting the demand for the rebar market.

Therefore, it is of great significance to model and predict the demand for reinforcement. How to fully grasp the characteristics of various indicators affected by the demand for reinforcement, as well as their relationships and relevant influencing factors, comprehensively evaluate and solve the long-term backward construction foundation, is benefit for improving the quality of new urbanization,

promoting effective investment, increasing the demand for reinforcement consumption, and creating a new driving force for economic development.

Recently, many scholars have carried out a large number of prediction models, and proposed time series, neural network, limit learning, machine learning, grey correlation analysis and other rebar demand prediction methods [1-3]. The neural network has the characteristics of parallelism, fault tolerance, hardware realization and self-learning, as well as good nonlinear fitting ability. The calculation method of the neural network is different from the traditional method. Wang Huaqiang et al. and Li Guoliang et al. proposed the IGA-BP network model by combining the global optimization of immune genetic algorithm and local optimization of BP network, Therefore, using neural network to establish the reinforcement demand prediction model is useful [4-6].

XGBOOST model was proposed by Dr. Chen Tianqi which is based on gradient Boosting integration algorithm [7]. And due to its high efficiency and accuracy of prediction results, XGBoot has been widely used in many aspects, such as face recognition in payment and login [8], fire recognition application [9], and precise train parking problem [10].

## 2. Data Description

### 2.1. Data Sources

The data used in this paper is from the open data set of the 2020 Dimension Cup International Mathematical Modelling Challenge for College Students, including the rebar apparent demand, and related variable data information from 2016 to 2020. In addition, the correlation coefficient between the apparent demand for rebar and 9 data indicators, such as construction area, new construction area, construction rebar quantity, rebar price, accumulative period of housing, national government fund income, national cement construction arithmetic average value are also included.

### 2.2. Handling of Outliers and Missing Values

We select the data package of eight factors affecting rebar demand for analysis, including rebar price, construction rebar trading volume, infrastructure investment, national cement construction start arithmetic average, national government fund income, screw purchase volume, housing area and commercial housing sales index [11].

According to the statistics of standardized mass sample data provided in the attachment, there are outliers and missing values in the analysis data. The existence of outliers often seriously affects the quality of modelling and prediction [12]. In order to reduce the impact of dimensional differences between relevant variables, the data will be preprocessed and standardized, so that it can improve the efficiency of comprehensive evaluation and comparison.

Based on the data from 2016 to 2020, the average value of relevant data in three months each year is taken as an indicator. In order to reduce the dimensional effect of differences between relevant variables, sample data were preprocessed and deviations were standardized. Through standardized line chart analysis, the efficiency of comprehensive evaluation analysis is improved.

### 2.3. Filter the Influencing Factors of Characteristics

The average of the relevant data in three months of each year is used as an indicator to preprocess the sample data and standardize the deviations in order to reduce the dimensional effect of the differences between the relevant variables. The test results using MATLAB are as follows:

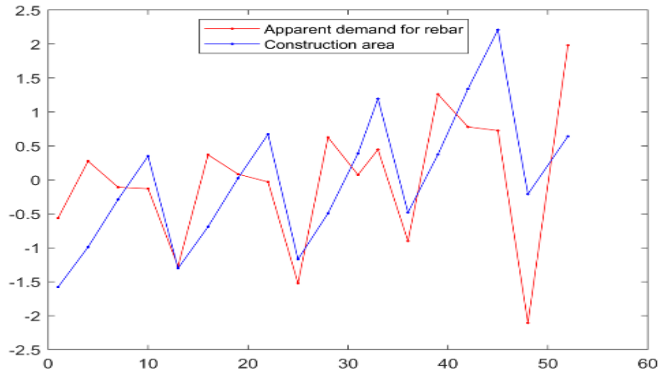


Figure 1: The relationship between apparent demand for rebar and Construction area.

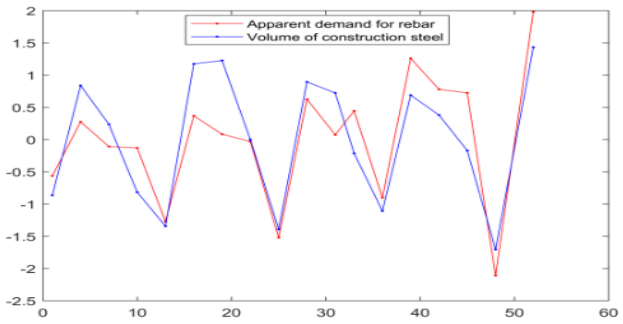


Figure 2: The relationship between apparent demand for rebar and Volume of Construction steel.


Figure 3: The relationship between apparent demand for rebar and Price of rebar.


Figure 4: The relationship between apparent demand for rebar and Thread purchase quantity.

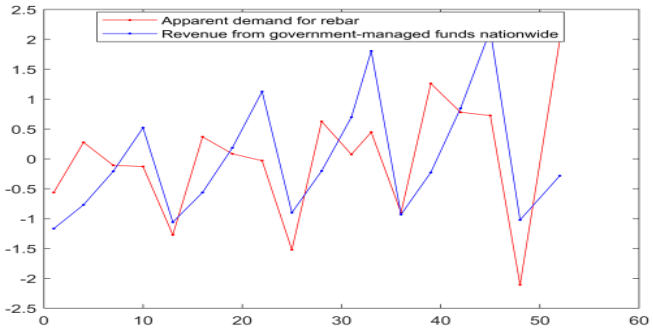


Figure 5: The relationship between apparent demand for rebar and Revenue from government-managed funds nationwide.

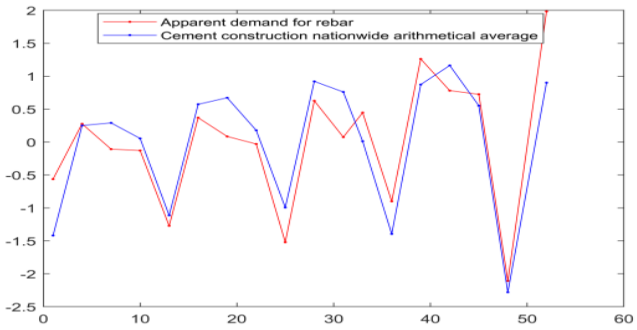


Figure 6: The relationship between apparent demand for rebar and Cement construction nationwide arithmetical average.

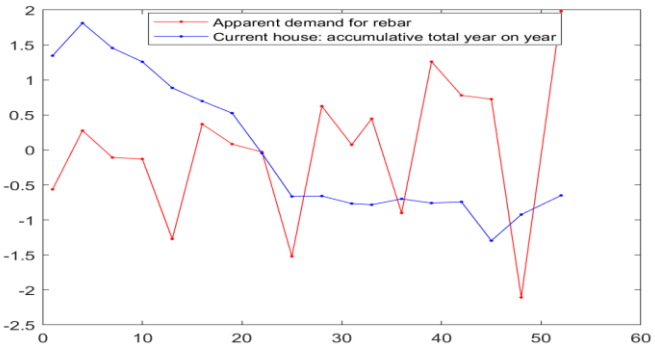


Figure 7: The relationship between apparent demand for rebar and current house (accumulative total year-on-year).


Figure 8: The relationship between apparent demand for rebar and Forward room (accumulative total year-on-year).

By the trend of the indicators in the Figure 1-8 above, it can be found that the demand for rebar has the same trend as the turnover of construction rebar, the arithmetic average of cement construction, the purchase quantity of wire screw, the area under construction and the area under new construction. Other factors are not consistent. It can be preliminarily find that the demand of rebar has a great correlation with the volume of construction rebar, the arithmetic average of cement construction, the purchase quantity of wire screw, the area of housing construction and the area of new housing construction.

## 3. Predictive Modelling

### 3.1. Pearson Correlation Coefficient Model

The Pearson correlation coefficient is defined as follows:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}\cdot\sqrt{n\sum_{i=1}^{n}y_i^2-(\sum_{i=1}^{n}y_i)^2}} \tag{1}$$

In Equation (1), r represents the correlation coefficient, n represents the number of samples, $x_i$ and $y_i$ representthe elements of the ith sample respectively.

When r=1, it means x and y are completely correlated and linearly related.

It take 1 for the degree of correlation coefficient takes 1 as standard. Then the correlation coefficient of each index is divided into three categories: when r>0.8 is called highly correlated; r<0.3 shows low correlation; 0.3<r<0.8 means the correlation was moderate.

### 3.2. XGBOOST Model

Since the Pearson correlation coefficient model can only express the correlation degree and cannot accurately express the error range, we adopted XGBOOST model for further analysis. This model was proposed by Dr. Chen Tianqi as the gradient promotion model. Its simplified expression is a model integrating multiple base decision trees. The decision tree is nonlinear fitting and can be quickly adjusted by pruning. The weight index of each leaf can be determined and the error can be analyzed.

Similar to CART tree, multiple weak regression trees are used to evaluate regions and predict values. Each tree is randomly sampled independently, which ensures that the data learned by each tree has different emphasis, and at the same time ensures the independence between trees. The model is as follows:

$$\hat{y}=\sum_{i=1}^{k}f_k(x_i),\ f_x\in F(F=\{f(x)\in\omega_{q(x)}\}) \tag{2}$$

In order to solve the regression problem, RMSE is used for the loss function:

$$loss=\sqrt{\sum_{i=1}^{n}(\hat{y}^{(i)}-y^{(i)})^2} \tag{3}$$

The objective function E1 can be written as:

$$E_1 = \sum_{i=1}^{n} (y_i - y_i^{(t)})^2 + \sum_{i}^{T} \Omega(f_i) \tag{4}$$

In Equation (4), the first part is the loss function, and the second part is the regularization term, where the regularization term is derived from the addition of the regularization of K trees, and F represents all regression trees.

In step t, $y_i$ and $y_i^{(t-1)}$ are known values, and the simplified objective function can beobtained:

$$E_2 = \sum_{i=1}^{n} (y_i - (y_i^{(t-1)} - f_t(x_i)))^2 + \sum_{i}^{T} \Omega(f_i) \tag{5}$$

Based on Equation (2), the final objective function E3 is [13-15]:

$$E_3 = \sum_{j=1}^{T} ((\sum_{i \in I_j} g_i)\omega_j + \frac{1}{2}(\sum_{i \in I_j} h_j + \lambda)\omega_j^{\,2}) + \gamma T \tag{6}$$

Where j is the j regression tree, $\omega$ is the weight of the regression tree. Ij essentially represents a set, each value in the set represents the ordinal number of a training sample, and the whole set is the training sample divided into the j leaf node by the T CART tree.

## 4. Results and Interpretation

According to Pearson correlation coefficient of apparent demand for rebar and Volume of construction rebar (North), as well as thread purchase quantity are all greater than 0.8, achieving a high correlation. The Pearson correlation coefficient between the demand and construction area, construction starts, as well as deposit and advance payment: accumulated value, and cumulative value are between 0.8 and 0.3, which reaches a moderate correlation. The Pearson correlation coefficient between the demand and national government management fund and income, as well as the national average long-term housing of cement construction (cumulative year-on-year) is less than 0.3, indicating low correlation.

The weight distribution is show by the following Figure 9:

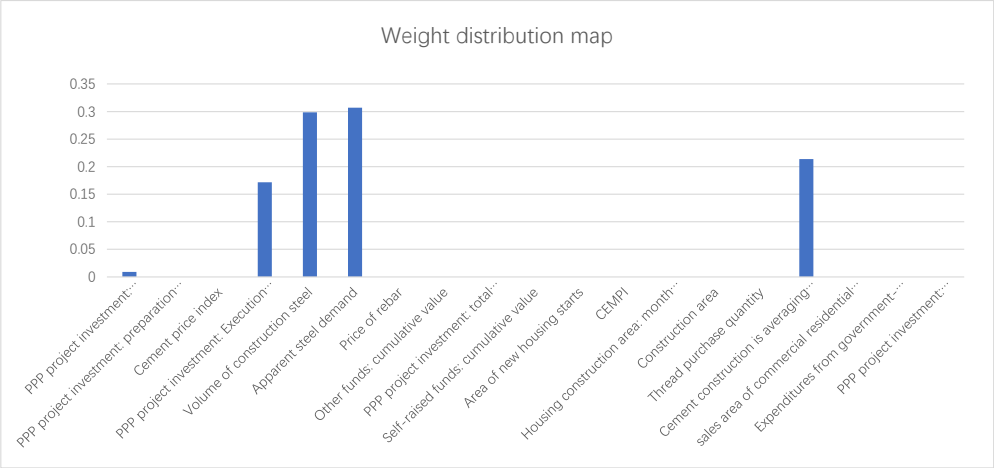


Figure 9: Weight distribution map of indicators related to the apparent demand of rebar.

The weight distribution diagram shows that the counterweight is mainly distributed in apparent

demand of rebar, national arithmetic average of cement construction, volume of construction rebar, PPP project investment: identification stage, PPP project investment: execution stage.

Based on the analysis, XGBoot training was used to test the mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE) and mean Absolute percentage error (MAPE) after 100 times of training. The results are shown in the Table 1:

Table 1: XGboost training error analysis.

| Evaluation index | Evaluation result |
|---|---|
| MSE | 21800.3782 |
| RMSE | 147.6495 |
| MAE | 256.8666 |
| $R^2$ | 0.8835 |
| MAPE | 16.8993 |

Therefore, we can see that Volume of construction rebar (North) is 0.8835, indicating that the regression relationship can explain the variation of the strain variable of 88.35%. Moreover, due to the large base of the test, RMSE is relatively small compared with the original sample data, so it can be considered that the test error of XGboost model is within a controllable range.

## 5. Conclusions

In this paper, the degree of correlation can be expressed by Pearson correlation coefficient of each influencing factor. Then, we use XGboost model to predict, in order to increase accuracy and persuasiveness, also add weight ratio analysis, decision tree error analysis, and enhance the reliability of prediction. Finally, it can be seen that the volume of construction rebar, apparent demand for rebar, and the weight index of the national arithmetic average are relatively high. In the Pearson correlation coefficient model of each influencing factor, all of them are moderately correlated, while other factors with large correlation coefficients account for less in the weight index.

Through the establishment of a line chart analysis between the apparent demand of rebar and various indicators, the macroscopic sieving out of 5 indicators consistent with the changing trend of rebar demand, and further use the Person correlation coefficient analysis to obtain the correlation coefficient of rebar procurement volume is greater than 0.8, and other indicators are greater than 0.3. In order to verify the reliability of this correlation, the above analysis accuracy is confirmed by the XGBOOST model and the embedded method in feature selection is used to screen the features, the weight ratio analysis and decision tree error analysis on different features are increased, and the return coefficient of the purchase volume of the visualized output steel reaches 0.8835, the MAME is 16.899%, and the importance of the features such as RMSE and MAE is within the controllable range. Therefore, in the analysis of rebar demand in the characteristic importance, the national rebar procurement volume contributes the most, and more attention should be paid to it. Some moderately relevant indicators, such as the amount of rebar reinforcement built and the national arithmetic average of cement construction, also need to be considered to increase the possibility of accurate predictions in the future.

## Acknowledgements

Research supported in part by Maoming science and technology project (190402001700308); Project of Education and Teaching Reform of Guangdong University of Petrochemical Technology (JY202030). Research supported in part by Innovation and Entrepreneurship Training Program for Undergraduate Students (73321127).

## References


*[1] Zhang G.Q., Patuwo B.E., Hu M.Y. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 1998, 14 (1): 35-62. [doi:10.1016/S0169.2070(97)00044-7].*

*[2] Martinez A., Schmuck C., Pereverzyev Jr. S., et al. A machine eaming framework for customer purchase prediction in the non-contractual setting. European Journal of Operational Research, 2020, 281 (3): 588-596. [doi:10.1016/j.ejor.2018.04.034].*

*[3] Zhou Xiaoxi, Xu Xing, Meng Jianfei, et al. Research progress of clothing sales forecasting methods. The needle 2020, (3), 68-72. [doi: 10.3969/j.issn.1000-4033.2020.03.017].*

*[4] Frank C., Garg A., Sztandera L.M., et al. Forecasting women's apparel sales using mathematical modelling. International Journal of Clothing Science and Technology, 2003, 15 (2): 107-125. [doi:10.1108/09556220310470097].*

*[5] Wang Huaqiang, Hu Ping, Li Haibo. Iga-bp network model for silicon content in molten iron of blast fur nace Application in quantity prediction. Journal of Hefei University of Technology (Natural Science) 2007, 3 0( 4): 413-415, 427.*

*[6] Li Guoliang, Li Zhongfu, Xie Hongtao, et al. Wavelet deities based on IGA-BP algorithm Network model and application. Systems Engineering, 2012, 30 (10): 112-117.*

*[7] CHENTQ, GUESTRINC. XGBoost: A Scalable Tree Boosting System. arXiv: 1603.02754 [cs.LG]. (2016-03-09). https://arxiv.org/abs/1603.02754.*

*[8] Du Xiaoxu, A Face Recognition Based on Boosting Algorithm. Hangzhou: Zhejiang University, 2006.*

*[9] Yang Guotian, Wu Zhangxian, Yang Pengyuan. Boosting application in Fire Recognition. Computer engineering and applications, 2010, 46 (5): pp.200-204.*

*[10] Zhou ji, Chen dewang. Application chine learning to train precise parking problem. The computer industry Process and applications, 2010, 46 (25): pp.226-230.*

*[11] 360 encyclopedia, deformed rebar bar, https://baike.so.com/doc/1212183-1282274.html, 2020-11-28.*

*[12] Zhang deran. Test method of outliers in statistical data. Statistical research, 2003, 20 (5): 53-55.*

*[13] Chen T., He T., Benesty M. xgboost: Extreme Gradient Boosting. 2016, 5 (9): 222-208.*

*[14] Wu Lichun. Research on Fault Diagnosis Algorithm based on Genetic Neural Network [Master degree]. Shenyang: Liaoning University, 2012.*

*[15] Chester Curme, Tobias Preis, H. Eugene Stanley, et al. Quantifying the semantics of search behavior before stock market moves. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111 (32): 11600-11605.*