Proceedings of Machine Learning Research vol 278:1–10, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

# A Hybrid XGBoost and Stacked Regression Model with Optimized Feature Selection for Port Throughput Prediction

**Ning Ding** DAISY_DN@OUTLOOK.COM
*International Business College, Dongbei University of Finance and Economics, Dalian, 116025, China*

**Fuyang Zhao*** ZHAOFUYANG1982@126.COM
*School of Economics and Management, Dalian Jiaotong University, Dalian, 116028, China*

**Xiaoyu Wang** 15941790913@163.COM
*School of Economics and Management, Dalian Jiaotong University, Dalian, 116028, China*
**Corresponding author*

**Editors:** Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

## Abstract

The prediction of port throughput is very important for port operation management. Aiming at the prediction of port throughput, this study selected the level of economic and trade and development vitality in the hinterland, regional transportation capacity in the hinterland, port infrastructure conditions and other first-class indicators and 13 second-class indicators, used xgboost to analyze the importance of characteristics, and screened out 9 key influencing factors. The combined model based on xgboost and stacking algorithm was constructed, and the parameters were optimized by cross validation and grid search method. Taking Dalian port as an example, the experiment shows that when xgboost stacking model is used to predict port throughput, MAE, MAPE and RMSE are the lowest, and $R^2$ is the highest. The prediction performance is significantly better than other models, which verifies the effectiveness and superiority of the model in port throughput prediction, and provides a new method and idea for port throughput prediction.

**Keywords:** Port throughput forecast; Xgboost algorithm; Stacking algorithm

## 1. Introduction

As the key node of logistics and transportation, the prediction of port throughput is very important for port operation management and resource allocation. In recent years, with the development of big data and machine learning technology, using advanced algorithms to accurately predict port throughput has become a research hotspot. This can not only provide a scientific basis for port planning, but also help to improve the operation efficiency and service quality of the port. However, there are many factors that affect the port throughput. How to accurately screen and use these factors to predict is a big challenge. In addition, the existing prediction models have their own advantages and disadvantages. How to combine the advantages of various models to build a more accurate prediction model is also worth discussing.

In order to improve the accuracy of port throughput prediction, the key link is to select the appropriate throughput influencing factors. Many scholars have conducted in-depth research on the influencing factors of port throughput from different angles (Eskafi et al., 2021; Tang et al., 2019; Li et al., 2017; Chen et al., 2018; Zhang et al., 2016). There are many prediction models for port cargo throughput, such as ARIMA model based on time series (Li, 2024) and improved GM prediction model (Wang and Phan, 2014); Support vector regression based on machine learning

(Xie et al., 2013); Stacking ensemble learning (Ma et al., 2024); Long and short term memory neural network (LSTM) (Wu and Lin, 2019), convolutional neural network (CNN), recurrent neural network (RNN) (Su et al., 2019) and Grey-CNN (Zeng and Xu, 2024) based on deep learning (Jiang et al., 2021); comprehensive forecasting model using VMD-CNN-GRU (Tan and Huang, 2024). In the field of port throughput prediction, previous studies provide a wealth of theoretical basis and practical experience. Some scholars have discussed the impact of port hinterland on port throughput from different angles, and proposed that socio-economic indicators be used as the input of the prediction model. Some scholars have integrated a variety of influencing factors, determined the main factors affecting port throughput through systematic clustering and principal component analysis, and analyzed the influencing factors of port throughput from the perspective of system, population and shipping development. However, the influencing factors of port throughput are numerous and complex, such as hinterland economy, Global trade situation, national policies and so on. How to accurately screen and use these factors to predict is the difficulty of research. At the same time, the existing research still has some limitations, such as the lack of comprehensive consideration of a variety of influencing factors and the limited generalization ability of the model. In recent years, machine learning algorithms have made remarkable achievements in the field of data prediction, and xgboost algorithm has attracted extensive attention because of its high efficiency and accuracy. Therefore, this study aims to build a more accurate port throughput prediction model by combining xgboost and stacking algorithm. This study provides a new method and idea for port throughput prediction.

## 2. Data Processing and Factor Selection

### 2.1. Data processing

According to the availability of port throughput data, the hinterland economic and trade level and development vitality, hinterland regional transportation capacity, and port infrastructure conditions are selected as the first level indicators, and 13 specific second level indicators are selected for measurement. Table 1 lists all influencing factors and their corresponding indicator codes. Table 2 and table 3 provide the specific data of the factors affecting the throughput of Dalian port. Table 2 mainly involves the data of economy and trade, while Table 3 includes the data of transportation and infrastructure.

### 2.2. Factor selection

Xgboost is used to analyze the correlation between 13 preselected indicators and port throughput. These 13 pre selected indicators cover many aspects, such as the economic and trade level and development vitality of the hinterland, the transportation capacity of the hinterland and the port infrastructure conditions. Through programming and data analysis using jupyter software based on python, the analysis results of feature importance are obtained, as shown in Figure 1.

Among all the 13 pre selected indicators, GDP, The total import and export volume, the total retail sales of consumer goods, the output value of the tertiary industry, the actual use of foreign capital, the total freight volume, the total freight turnover volume, the waterway freight turnover volume and the length of coastal terminals have the most significant impact on the port throughput of Dalian Port. These indicators occupy a relatively high position in the ranking, indicating that they

Table 1: Factors affecting port throughput

| Level 1 indicators | Specific indicators | Indicator code |
|---|---|---|
| Level of economic trade and development dynamism in the hinterland | GDP | X1 |
| | Per capita GDP | X2 |
| | Total exports and imports | X3 |
| | Total retail sales of consumer goods | X4 |
| | Primary sector output | X5 |
| | Secondary sector output | X6 |
| | Tertiary output | X7 |
| | Actual utilization of foreign capital | X8 |
| Hinterland regional transport capacity | Total freight volume | X9 |
| | Total freight turnover | X10 |
| | Waterborne freight turnover | X11 |
| Port infrastructure conditions | Length of coastal piers | X12 |
| | Production 10,000 tons berths | X13 |

Table 2: Part of the map of factors affecting throughput at Dalian port (1)

| Vintages | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|---|---|---|---|---|---|---|---|
| 2015 | 5181.9 | 74181 | 645.9 | 1602.5 | 333.7 | 2483.2 | 2365 |
| 2016 | 5400.1 | 76457 | 550.9 | 1692.4 | 381.2 | 2245.3 | 2773.6 |
| 2017 | 5648 | 79125 | 514.7 | 1827 | 404.1 | 2289.8 | 2954.1 |
| 2018 | 6052.2 | 83988 | 4132.2 | 1955.6 | 412.9 | 2428.5 | 3210.8 |
| 2019 | 6500.9 | 89255 | 4701.4 | 2043 | 430 | 2477.9 | 3593 |
| 2020 | 6990 | 94966 | 4352.8 | 2064.5 | 458.9 | 2806.8 | 3724.3 |
| 2021 | 7000.4 | 94281 | 3852.9 | 1828 | 459.2 | 2804.1 | 3737.1 |
| 2022 | 7825.9 | 104751 | 4248.5 | 1909.7 | 513.3 | 3301.6 | 4011 |

Table 3: Part of the map of factors affecting throughput in Dalian harbour (2)

| Vintages | X8 | X9 | X10 | X11 | X12 | X13 |
|---|---|---|---|---|---|---|
| 2015 | 140 | 44736 | 83357742 | 77830171 | 39449 | 98 |
| 2016 | 27 | 42002 | 83136624 | 78263012 | 43956 | 103 |
| 2017 | 30 | 43116 | 86382299 | 81372073 | 40765 | 103 |
| 2018 | 32.5 | 44955 | 90074513 | 84620023 | 41101 | 104 |
| 2019 | 26.8 | 46570 | 68098613 | 62206179 | 41101 | 104 |
| 2020 | 8.7 | 34158 | 54116364 | 49382000 | 41101 | 104 |
| 2021 | 6.6 | 25877 | 19763235 | 15073992 | 43218 | 111 |
| 2022 | 16.7 | 22811 | 7926080 | 4882257 | 43268 | 109 |

have a greater weight on the prediction and impact of port throughput. The final index screening results are shown in Table 4.

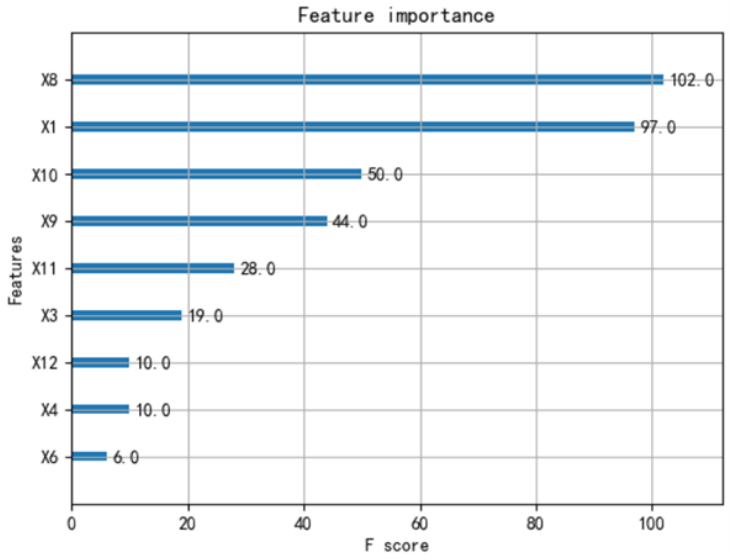


Figure 1: Characteristic importance map

Table 4: Screened port throughput influencing factors

| Level 1 indicators | Specific indicators | Indicator code |
|---|---|---|
| Level of economic trade and development dynamism in the hinterland | GDP | X1 |
| | Total exports and imports | X3 |
| | Total retail sales of consumer goods | X4 |
| | Tertiary output | X7 |
| | Actual utilization of foreign capital | X8 |
| Hinterland regional transport capacity | Total freight volume | X9 |
| | Total freight turnover | X10 |
| | Waterborne freight turnover | X11 |
| Port infrastructure conditions | Length of coastal piers | X12 |

## 3. Model Construction and Parameter Optimization

### 3.1. Mathematical Model

In the prediction of port throughput, multiple linear regression model and SVR algorithm can be used to analyze and predict port throughput. Introducing relaxation variables $\xi_i$ and $\xi_i^*$, the objective function of SVR is shown in formula (1).

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{\mathrm{m}}(\xi_i + \xi_i^*) \tag{1}$$

$$s.t. \begin{cases} y_i - \omega x - b \leq \in + \xi_i, \xi_i \geq 0 \\ \omega x + b - y_i \leq \in + \xi_i^*, \xi_i^* \geq 0 \end{cases}$$

In order to minimize the objective function, a Lagrange function is constructed according to the constraints, as shown in formula (2).

$$\begin{aligned} L &= \frac{1}{2}\|\omega\|^2 + C\sum\nolimits_{i=1}^{m}\left(\xi_{\mathrm{i}} + \xi_i^*\right) - \sum_{i=1}^{m}\alpha_i\left(\in + \xi_i - y_i + \omega x + b\right) \\ &- \sum_{i=1}^{m}\alpha_i^*\left(\in + \xi_i^* + y_i - \omega x - b\right) - \sum_{i=1}^{m}\left(\eta_i\xi_i + \eta_i^*\xi_i^*\right) \\ s.t.&\quad \alpha_i,\alpha_i^*,\eta_i,\eta_i^* \geq 0 \end{aligned} \tag{2}$$

Let the partial derivative of to $w$, $b$,$\xi_i$, $\xi_i^*$ be zero, and the formula (3) can be obtained.

$$\omega = \sum_{i=1}^{m}\left(\alpha_i^* - \alpha_i\right)x_i \tag{3}$$

The above process needs to meet the KKT condition, and the kernel function is introduced. Assuming that the optimal solution of the dual problem is $(a_1, a_1^*, a_2, a_2^*, ....., a_n, a_n^*)$, the final support vector regression model is formula (4).

$$f\left(x\right) = \sum_{i=1}^{m}\left(a_i - a_i^*\right)K\left(x_i x\right) + b \tag{4}$$

The objective function of xgboost model can be divided into error function term $L$ and model complexity function term $\Omega$. The objective function is shown in formula (5).

$$OBJ = L + \Omega \tag{5}$$

$$L = \sum\nolimits_{i=1}^{n}\left(y_i - \underset{i}{\wedge}{y}\right)^2 \tag{6}$$

$$\Omega = \gamma T + \frac{1}{2}\lambda\sum_{\mathrm{j}=1}^{T}\omega_j^2 \tag{7}$$

When using the training data to optimize the model, it is necessary to keep the original model unchanged and add a new function $f\left(x\right)$ to the model to reduce the objective function as much as possible. At this time, the objective function is expressed as formula (8).

$$Obj^{(t)} = \sum_{i=1}^{n}\left(y_i - \left(\overset{\wedge}{y}^{(t-1)} + f_i(x_i)\right)\right)^2 + \Omega \tag{8}$$

In Xgboost algorithm, in order to quickly find the parameters that minimize the objective function, the objective function is expanded by second-order Taylor expansion, and the approximate objective function is obtained as shown in formula (9).

$$Obj^{(t)} \approx \sum_{i=1}^{n}\left(\left(y_i - \overset{\wedge}{y}^{(t-1)}\right)^2 + 2\left(y_i - \overset{\wedge}{y}^{t-1}\right)f_t(x_i) - h_i f_t^2(x_i)\right) + \Omega \tag{9}$$

When the constant term is removed, it can be seen that the objective function is only related to the first and second derivatives of the error function. At this time, the objective function is expressed as formula (10).

$$Obj \approx \sum_{i=1}^{n}\left[g_i\omega_{q(x_i)}+\frac{1}{2}h_i\omega_q^2\left(x_i\right)\right]+\gamma T+\frac{1}{2}\sum_{j=1}^{T}\omega_j^2=\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j}g_i\right)\omega_j+\frac{1}{2}\left(\sum_{i\in I_j}h_i+\lambda\right)\omega_j^2\right]+\gamma T \quad (10)$$

If the structural part $q$ of the tree is known, the objective function can be used to find the optimal $\omega_j$ and obtain the optimal objective function value. Its essence can be classified as the problem of solving the minimum value of quadratic function. The solution is obtained as shown in formula (11) and formula (12).

$$\omega_j^*=\frac{-\sum_{i\in I_j}g_i}{\sum_{i\in I_j}h_i+\lambda} \quad (11)$$

$$Obj=-\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j}g_i\right)^2}{\sum_{i\in I_j}h_i+\lambda}+\gamma T \quad (12)$$

GBDT algorithm can be regarded as an additive model composed of trees, and its corresponding formula is shown in (13).

$$F\left(x,w\right)=\sum_{m=0}^{M}\alpha_m h_m\left(x,w_m\right)=\sum_{m=0}^{M}f_m\left(x,w_m\right) \quad (13)$$

Where $x$ is the input sample; $w$ is the model parameter; $h$ is classified regression tree; $\alpha$ is the weight of each tree. The implementation process of GBDT algorithm is as follows: given the training data set: $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$. Where, $x_i\in X\subseteq R^n$, $X$ is the input space, $y_i\in Y\subseteq R$, $Y$ is the output space, and the loss function is $L\left(y,f\left(x\right)\right)$. The goal is to get the final regression tree $F_M$.

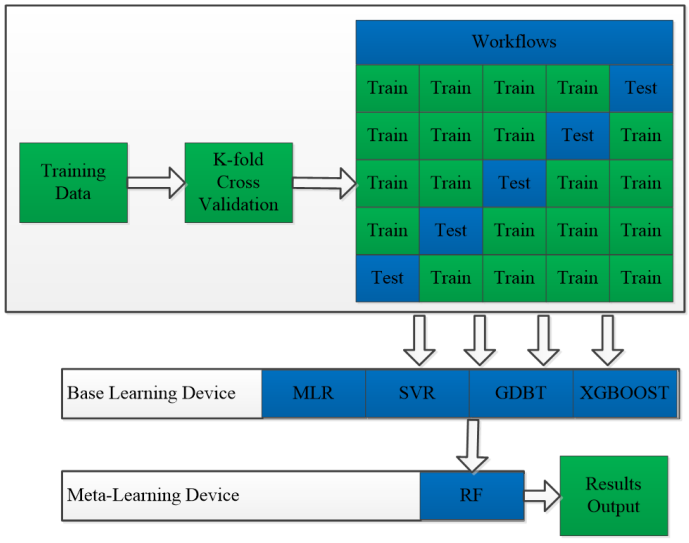


Figure 2: Schematic diagram of the Stacking "two-phase" model

In order to improve the prediction accuracy of port throughput, stacking algorithm is applied as an advanced machine learning method. The algorithm uses a two-tier structure. In the first layer, feature extraction and information extraction are performed on the data through a number of different base learners. During the training process, these basic learning devices dig deep into the data set and effectively extract the features related to port throughput; In order to solve the over fitting problem, the random forest ensemble learning method is introduced as the second layer classifier. Combining the two-tier structure of stacking algorithm and the advantages of random forest, a robust prediction model can be built to accurately predict the port throughput. The "two-stage" model of stacking is shown in Figure 2.

### 3.2. Parameter Optimization

The method of cross validation and grid search is used to determine the superparameters of the model. The specific process is shown in Figure 3. This paper adopts the strategy of fixing most parameters and tuning only a few key parameters. Through cross validation, a set of superparameter combinations with the highest average score are found as the optimal parameters of the model. This method aims to improve efficiency and ensure the best prediction performance of the found parameter combination.

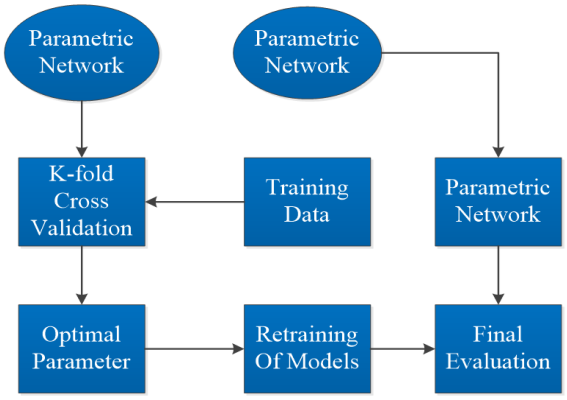


Figure 3: Flow chart of grid search method

## 4. Prediction Process and Result Comparison

### 4.1. Prediction Process

The specific process of Port Throughput Prediction Research Based on xgboost algorithm and stacking combination model is shown in Figure 4.

### 4.2. Result Comparison

In this comparative experiment, the port throughput data of Dalian port from 1999 to 2022 are selected as samples. In this experiment, four basic learning machine models are used: multiple linear regression, support vector regression (SVR), gradient lifting decision tree (GBDT) and extreme gradient lifting (xgboost). In order to find the optimal parameters of each model, a method combining

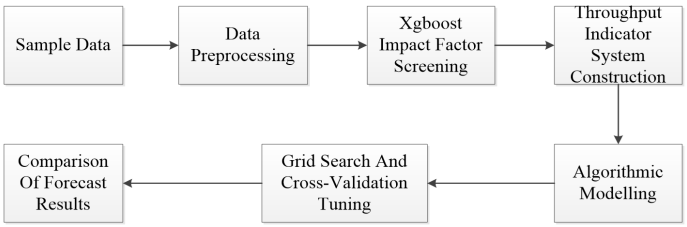


Figure 4: Flow chart of port throughput forecasting

grid search and cross validation is used. In Table 5, we can see the performance comparison results of the five models in predicting the port throughput of Dalian port.

Table 5: Comparison of model data

| data set | MLR | SVR | GBDT | XGBoost | XGBoost-Stacking |
|---|---|---|---|---|---|
| MAE | 2979 | 4107 | 1563 | 1870 | 650 |
| MAPE | 11% | 13% | 6% | 7% | 3% |
| RMSE | 4642 | 6558 | 1990 | 2211 | 969 |
| $R^2$ | 82.8% | 65.8% | 96.8% | 96.1% | 99.2% |

According to the comparative data in Table 5, it can be confirmed that xgboost stacking model shows significant advantages in port throughput prediction. This is because xgboost algorithm can effectively deal with the interaction between features, while stacking algorithm can improve the overall prediction performance by integrating the prediction results of multiple base learners. For these two reasons, xgboost stacking model can better adapt to the complex data structure and characteristics, thus showing excellent performance in predicting port throughput.

According to figures 5 and 6, it can be observed that the stacking model performs better than other models at each prediction time node. From the overall trend, the range of predictive values of stacking model is relatively stable, while the range of predictive values of other models fluctuates greatly. This further proves the robustness of the stacking model at different prediction time nodes.

## 5. Conclusion

This study focuses on predicting port throughput, selecting primary indicators such as hinterland economic and trade level and vitality, regional transportation capacity, and port infrastructure, as well as 13 secondary indicators. Using Xgboost to analyze feature importance, 9 key influencing factors were selected. Build a combined model based on Xgboost and Stacking algorithm, and optimize parameters through cross validation and grid search method. Experiments have shown that the Xgboost Stacking model has the lowest MAE, MAPE, and RMSE, and the highest $R^2$ when predicting the throughput of Dalian Port. Its predictive performance is significantly better than other models, verifying its effectiveness and superiority. However, there may be potential factors that have not been considered in the screening of influencing factors; In model construction, there is subjectivity in parameter setting and adjustment, which may affect the prediction results. Future research can expand the sample size, include more potential influencing factors, further optimize model parameters, and improve model generalization ability and prediction accuracy.

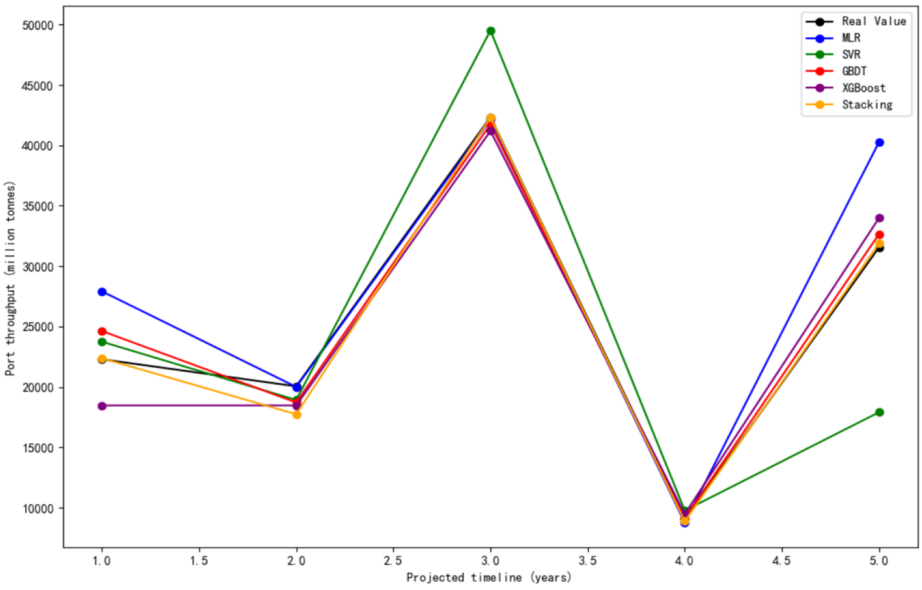


Figure 5: Line graph for model comparison

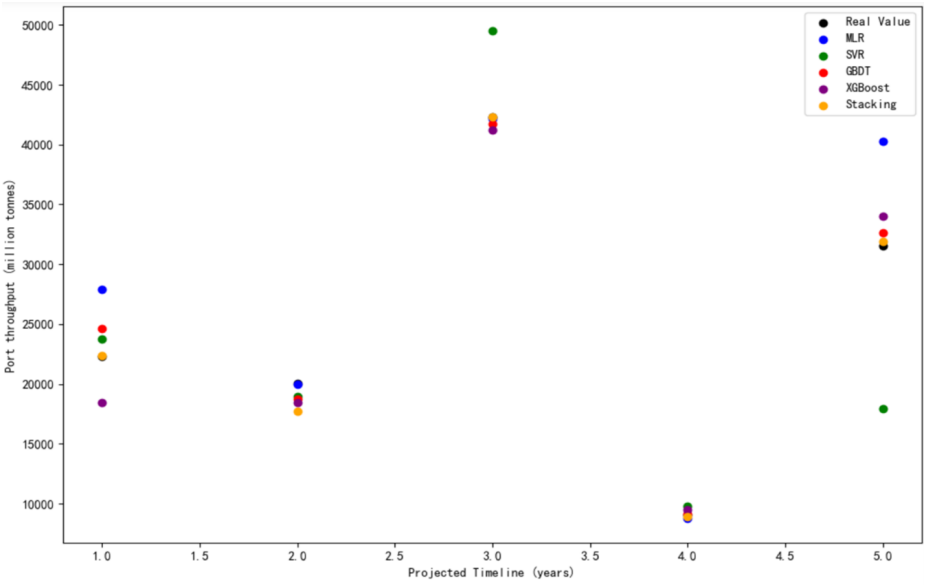


Figure 6: Scatter plot for model comparison

## References

J. Chen, Y. Fei, Y. Zhu, and et al. Allometric relationship between port throughput growth and urban population: A case study of shanghai port and shanghai city. *Advances in Mechanical Engineering*, 10(3):168–183, 2018. doi: 10.1177/1687814018760933.

M. Eskafi, M. Kowsari, A. Dastgheib, and et al. Mutual information analysis of the factors influencing port throughput. *Maritime Business Review*, 6(2):129–146, 2021. doi: 10.1108/MABR-05-2020-0030.

F. Jiang, G. Xie, and S. Wang. Forecasting port container throughput with deep learning approach. In *Proceedings of the 5th International Conference on Computer Science and Application Engineering*, pages 123–128, 2021. doi: 10.1145/3487075.3487173.

J. Y. Li, T. E. Notteboom, and J. J. Wang. An institutional analysis of the evolution of inland waterway transport and inland ports on the pear river. *GeoJournal*, 82(5):867–886, 2017. doi: 10.1007/s10708-016-9696-0.

M. Li. Port throughput forecast based on arima model-take tianjin port as an example. *Highlights in Science, Engineering and Technology*, 105:11–17, 2024. doi: 10.54097/dv0kzs60.

M. Ma, Z. Sun, P. Han, and H. Yang. A stacking ensemble learning for ship fuel consumption prediction under cross-training. *Journal of Mechanical Science and Technology*, 38(1):299–308, 2024. doi: 10.1007/s12206-023-1224-9.

H. Su, E. Zio, J. Zhang, and et al. A hybrid hourly natural gas demand forecasting method based on the integration of wavelet transform and enhanced deep-rnn model. *Energy*, 178:585–597, 2019. doi: 10.1016/j.energy.2019.04.167.

Q. Tan and H. Huang. Comprehensive forecasting model for port container throughput based on hybrid deep neural networks. In *Proceedings of the 2024 7th International Symposium on Traffic Transportation and Civil Architecture*, volume 241, pages 332–341, 2024. doi: 10.2991/978-94-6463-514-0_35.

S. Tang, S. Xu, and J. Gao. An optimal multifactor-based model for container throughput forecasting. *KSCE Journal of Civil Engineering*, 23(9):4124–4131, 2019. doi: 10.1007/s12205-019-2446-3.

C.-N. Wang and V.-T. Phan. An improvement the accuracy of grey forecasting model for cargo throughput in international commercial ports of kaohsiung. *International Journal of Business and Economic Research*, 3(1):1–5, 2014. doi: 10.11648/j.ijber.20140301.11.

Q. Wu and H. Lin. Daily urban air quality index forecasting based on variational mode decomposition, sample entropy and lstm neural network. *Sustainable Cities and Society*, 50:1125–1141, 2019. doi: 10.1016/j.scs.2019.101657.

G. Xie, S. Wang, Y. Zhao, and et al. Hybrid approaches based on lssvr model for container throughput forecasting: A comparative study. *Applied Soft Computing Journal*, 13(5):2232–2241, 2013. doi: 10.1016/j.asoc.2013.02.002.

F. Zeng and S. Xu. A hybrid container throughput forecasting approach using bi-directional hinterland data of port. *Scientific Reports*, 14(1):25502, 2024. doi: 10.1038/s41598-024-77376-9.

Y. Zhang, J. Siu, and L. Lam. Estimating economic losses of industry clusters due to port disruptions. *Transportation Research Part A*, 91:11–33, 2016. doi: 10.1016/j.tra.2016.05.017.