Forecasting the Busan Container Volume Using XGBoost Approach based on Machine Learning Model

Nguyen Thi Phuong Thanh[1], Gyu Sung Cho[2*]
[1]Student, Department of Port Logistics System, Tongmyong University
[2]Professor, Department of Port Logistics System, Tongmyong University

기계 학습 모델을 통해 XGBoost 기법을 활용한 부산 컨테이너 물동량 예측

웬티프엉타인[1], 조규성[2*]
[1]동명대학교 항만물류시스템학과 학생, [2]동명대학교 항만물류시스템학과 교수

Abstract Container volume is a very important factor in accurate evaluation of port performance, and accurate prediction of effective port development and operation strategies is essential. However, it is difficult to improve the accuracy of container volume prediction due to rapid changes in the marine industry. To solve this problem, it is necessary to analyze the impact on port performance using the Internet of Things (IoT) and apply it to improve the competitiveness and efficiency of Busan Port. Therefore, this study aims to develop a prediction model for predicting the future container volume of Busan Port, and through this, focuses on improving port productivity and making improved decision-making by port management agencies. In order to predict port container volume, this study introduced the Extreme Gradient Boosting (XGBoost) technique of a machine learning model. XGBoost stands out of its higher accuracy, faster learning and prediction than other algorithms, preventing overfitting, along with providing Feature Importance. Especially, XGBoost can be used directly for regression predictive modelling, which helps improve the accuracy of the volume prediction model presented in previous studies. Through this, this study can accurately and reliably predict container volume by the proposed method with a 4.3% MAPE (Mean absolute percentage error) value, highlighting its high forecasting accuracy. It is believed that the accuracy of Busan container volume can be increased through the methodology presented in this study.

Key Words : Internet of Things (IoT), container throughput forecasting, cargo volume, machine learning (ML), XGBoost technique, Busan Port

요 약 항만 성능에 대한 정확한 평가는 컨테이너 물동량은 매우 중요한 요소이며, 효과적인 항만 개발 및 운영 전략에 대한 정확한 예측이 필수적이다. 하지만 해양 산업의 급격한 변화로 인해 컨테이너 물동량 예측의 정확성이 향상되기는 어렵다. 이를 해결하기 위해 사물인터넷(IoT)을 이용한 항만 성능에 미치는 영향을 분석하여 부산항의 경쟁력과 효율성을 향상시키기 위해 적용이 필요하다. 이에 본 연구에서는 부산항의 미래 컨테이너 물동량을 예측하기 위한 예측 모델을 개발하는 것을 목표로 이를 통해 항만 관리 기관의 개선된 의사 결정과 항만 생산성을 향상시키는 데 초점을 맞추고 있다. 항만 컨테이너 물동량을 예측하기 위해 본 연구에서는 기계 학습 모델의 Extreme Gradient Boosting (XGBoost) 기법을 도입하였다. XGBoost는 다른 알고리즘에 비해 높은 정확도, 빠른 학습 및 예측 속도,과적합을 방지하고 Feature Importance 제공하는 장점이 돋보인다. 특히 XGBoost는 회귀 예측 모델링에 직접 사용할 수 있어 기존 연구에서 제시된 물동량 예측 모델의 정확도 향상에 도움이 된다. 이를 통해 본 연구는 4.3% MAPE (Mean absolute percenture error) 값으로 제안된 방법이 컨테이너 물동량을 정확하고 신뢰성 있게 예측할 수 있다. 본 연구에서 제시한 방법론을 통해서 부산 컨테이너물동량의 정확성을 높일 수 있을 것으로 판단된다.

주제어 : 사물인터넷 (IoT), 컨테이너 물동량 예측, 물동량, 기계학습 (ML), XGBoost 기법, 부산항

This research was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2023RIS-007).
*교신저자 : 조규성(gscho@tu.ac.kr)
접수일 2024년 01월 22일 수정일 2024년 02월 10일 심사완료일 2024년 02월 12일

1. Introduction

The maritime and port logistics industry is closely intertwined with global trade and economic activity, particularly for Korea, which heavily relies on trade. Busan Port, as the largest port in Korea and responsible for processing 75% of the country's container cargo, has attracted the attention of researchers interested in utilizing deep learning and machine learning models for time-series prediction [1-2]. The accurate prediction of container volume at Busan Port holds great importance for national competitiveness, and enhancing the accuracy of prediction models is essential for this purpose.

With breakthroughs in the field of artificial intelligence, the trend of utilizing machine learning approaches to predict has received much attention due to their robustness in dealing with large and complex data using Internet of Things(IoT)[3-5]. Machine learning enables computers to learn from data by identifying patterns and relationships in training data, allowing them to make forecasts for future values and events. A XGBoost-based machine learning strategy for estimating container throughput at Busan Port is introduced in this research. Its ability to handle missing data, highly parallelizable code, and robustness for both classification and regression problems make it a preferred choice for many data scientists and researchers. The rest of this paper is structured as follows. Section 2 provides a summary of existing research on container volume prediction and examines the study's contributions. Section 3 offers an overview of machine learning model and the XGBoost technique. Section 4 highlights the procedure, including data analysis, and evaluates the related error values for the suggested model. Finally, Section 5 presents the study's conclusion and future research.

2. Literature Review

Previous studies on container volume prediction have implemented traditional models such as autoregressive integrated moving average (ARIMA), seasonal autoregressive integrated moving average (SARIMA), and traditional regression. Yi, G. (2013) used a range of seasonal multiplicative ARIMA models to estimate and anticipate the container throughput of the Busan port. Based on AIC, SC, and Hannan-Quin information criterion, the model $(1,0,1)\times(1,0,1)12$ is the best. The chosen model's predicting results showed that the Busan port's container throughput would steadily rise annually between 2013 and 2020, with monthly volatility changes brought on by seasonality and other factors [6]. These methods, however, do not account for the complexity and variability of data caused by external factors such as the global financial crisis and economic volatility because the performance reduces when unexpected variations are reflected. In other words, various external factors can impact the stability of time-series data. Besides, a recent study by Lee, E., Kim, D., & Bae, H. (2021) proposed a novel deep learning prediction model that combines prediction and trend identification of container volume. By merging deep learning-based multivariate long short-term memory (LSTM) prediction with port volume time-series decomposition and external factors, the proposed model investigates external variables that are associated with container volume. The findings show that the suggested model outperforms the traditional LSTM model while also following the trend[7-8]. Doo-Hwan, K., & Lee, K. (2020) compared the performance of the LSTM model with the SARIMA model to improve the accuracy of container cargo volume forecasting. The results reveal that the LSTM model surpassed the SARIMA model in forecasting accuracy [6]. The traditional time-series prediction techniques benefit from fitting time series data and being simple and easy to understand. The trend and seasonality of time series data can be well

predicted. However, they have limited ability to handle nonlinearity in data and capture complex patterns, depend on the parameter setting, and may have low predictive power. On the other hand, XGBoost provides high predictive performance through ensemble learning using boosting techniques. XGBoost is resistant to overfitting and can be applied to different types of data. Moreover, XGBoost computes Feature Importance to facilitate model interpretation. Therefore, XGBoost has the advantage of being more complex, better performance, and applicable to different datasets than the traditional methods. According to Filom, S., Amiri, A. M., & Razavi, S. (2022), the number of studies in the field has been increasing year after year, and the most common application of machine learning approaches is to forecast particular port characteristics. Added to that, while the implementation of machine learning approaches in the maritime transportation and port industry has steadily attracted researchers, only a few studies have offered to review, categorize, and find research paths in the relevant. Barua et al. (2020) examined ML-based applications in international freight transport, with a particular emphasis on the shipping industry. Munim et al. (2020) explored big data and artificial intelligence in the maritime business[9].

3. Methodology

3.1 Overview of Machine Learning methods

Machine learning techniques are versatile tools that have the ability to uncover the untapped potential of data in critical areas such as ports within supply chain and transportation networks. These techniques involve the creation of machine learning models from algorithms, which are trained using labeled, unlabeled, or mixed data. Different algorithms are suitable for different objectives, such as classification or prediction modelling, and are adapted to effectively handle specific tasks, thereby becoming machine learning models. Machine learning leverages data to generate informed predictions and decisions without the need for prior knowledge of the data or context. The primary aim of machine learning forecasting is to enhance prediction accuracy while minimizing a loss function, aligning with traditional approaches. This article discusses the use of supervised learning algorithms, which rely on labeled datasets to train algorithms for accurate data classification and outcome prediction, where the value of the outcome of interest is known in historical or training data.

3.2 Overview of XGBoost technique

This paper introduces XGBoost as an implementation of the Gradient Boosted Decision Trees algorithm, chosen for its capacity to generalize other models by optimizing an arbitrary differential loss function. The core principle of the XGBoost algorithm is to minimize a specific objective function comprising the loss function and regularization terms [10].

$$Loss_{regression} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i}) \quad (1)$$

$$\Omega(f) = \gamma T + \lambda \parallel \omega \parallel^2 \quad (2)$$

$$Objective = Loss_{regression} + \Omega(f) \quad (3)$$

Here, n is the number of data points, y_i represents the actual target value for the i-th data point, and $\hat{y_i}$ denotes the predicted value by the model. Equation 1 (loss function) quantifies the discrepancy between the predicted values and the actual target values to minimize this loss during training. In equation 2 (regularization terms), T indicates the number of leaves in the tree, ω is a vector containing each leaf's score, f represents the combined output of all individual trees in the ensemble, and γ and λ are penalty coefficient. Regularization helps prevent overfitting by penalizing complex models. In other words, the goal is to figure out

the optimal combination of trees that minimizes the sum of the loss function and this regularization term. XGBoost algorithm combines the loss function and the regularization terms that is shown in equation 3 with an aim of finding the best set of trees that collectively minimize the overall error while maintaining simplicity. More explanations of the XGBoost technique can be found in Chen, T., & Guestrin, C. (2016). The process of building ensemble models involves creating new models and combining them into an ensemble. In each cycle, errors are calculated for each observation in the dataset, and a new model is built to predict those errors. The predictions from this error-predicting model are added to the ensemble, and the predictions from all previous models are used to make a decision. These predictions are then used to calculate new errors, build the next model, and add it to the ensemble. The cycle also requires an initial base prediction to start, which can be quite naive in practice. However, subsequent additions to the ensemble will address those errors[12-13].

3.3 The study's procedure

The procedure employing the XGBoost-based machine learning approach can be described as follows:

- Exploratory data analysis is conducted to examine the dataset's patterns through visualizing the data.
- During data preprocessing, missing values and outliers are removed.
- The dataset is divided into training and test sets after the date 01-01-2021.
- Time series features is created to visualize feature/target relationship.
- The XGBoost model is constructed with hyperparameters that have been optimized.
- The XGBoost model is evaluated by comparing the predicted results to the actual results from the test dataset through error metrics test set.
- The worst and best prediction was employed.

4. Data Analysis

The container volume was stored monthly beginning in March 2003 and lasted until November 2023. Train and test data are divided into about 80% (January 2013 ~ January 2021) and 20% (February 2021 ~ November 2023). All variables are numeric data, and the summary is shown in Table 1, which extracts part of the data due to big data.

〈Table 1〉 Dataset [12]

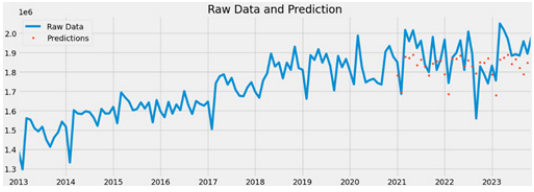
Month/Year	Total Container Throughput
01/01/2013	1391523.25
01/02/2013	1297834.25
01/03/2013	1561235.25
01/04/2013	1554039.25

In theory, XGBoost would implement the Regression model based on the singular or multiple features to predict future numerical values[13]. That is why the data training must also be in the numerical values. One easy way to get many new features and perspectives from which to view the data is to break down the datetime index into its many parts. The numerical features from the date enables to see in the Table 2. The table helps get a look at the vast array of new features extracted from one series of dates. This helps the model find more patterns within our data and leads to better accuracy for our predictions.

Fitting the data into the model is proceeded as soon as the required information is now all available. After the training process, having the prediction on the test data and visualizing them is also introduced in the graph. As can be seen, the prediction might seem slightly off but still follow the overall trend.

〈Table 2〉 The numerical features (extracted from the 5 first line)

Month/Year	Total Container Throughput	Quarter	Month	Year
01/01/2013	1391523.25	1	1	2013
01/02/2013	1297834.25	1	2	2013
01/03/2013	1561235.25	1	3	2013
01/04/2013	1554039.25	2	4	2013
01/05/2013	1493535.50	2	5	2013



[Fig. 1] The visualization of raw data and predicted data

Performance of machine learning models needs to be evaluated. To do that, the output generated by the models is compared with the actual values of the cargo volume, this is done by using the following performance indice: The mean absolute percentage error (MAPE), which is a measure of prediction accuracy of a forecasting method in statistics. It usually expresses the accuracy as a ratio defined by the formula[14]:

$$MAPE = \frac{1}{n} \times \Sigma \frac{|actual - forecast|}{|actual|} \times 100 \ (4)$$

The evaluation of the model using error metrics reveals that the prediction may have an error of approximately 4.3%. This means that the average difference between the forecasted value and the actual value is 4.3%, indicating highly accurate forecasting. Meanwhile, MAPE values of other models such as SARIMA, Univariate LSTM, and Time-series Decomposition is 22.14%, 18.10%, and 10.95% respectively. The MAPE evaluation metrics show that the error of the XGBoost model is the smallest, we can see that our method can bring the most effective forecasting performance among various prediction methods[15].

XGBoost also indicates the errors of the worst and best absolute predicted months. We can identify a few critical months where the predictions were notably inaccurate. For instance, as per the Chief Economist's Outlook in September 2022, the global economy has darkened due to high inflation and a global recession since mid-September 2022. Similarly, we can examine the dates when we achieved the best results. These dates do not

〈Table 3〉 The worst absolute predicted months (extracted from the 5 first line)

Year	Month	Total Container Throughput	Prediction	Error	Abs-Error
2022	9	1559996.75	1792082.875	-232086.125	232086.125
2023	3	2020054.50	1863988.750	186065.750	186065.750
2022	1	1966759.00	1788981.625	177777.375	177777.375
2023	9	1958643.25	1788017.625	170625.625	170625.625
2022	7	2008719.50	1859968.625	148750.875	148750.875

〈Table 4〉 The best absolute predicted months (extracted from the 5 first line)

Year	Month	Total Container Throughput	Prediction	Error	Abs-Error
2021	12	1864930.00	1858698.125	6231.875	6231.875
2022	3	1875668.00	1869121.000	6547.000	6547.000
2021	2	1686830.75	1693990.125	-7159.375	7159.375
2021	8	1838142.50	1826683.375	11459.125	11459.125
2021	9	1798941.00	1783523.250	17417.750	17417.750

appear to be clustered in any specific pattern, which is understandable given the model's high level of accuracy. The accurate predictions are distributed over a much wider range of time stamps.

5. Conclusion and Future research

In this research, the XGBoost machine learning algorithm was used to improve the accuracy of container volume prediction. The study found that the algorithm significantly improved prediction performance and trend prediction, which had a positive impact. However, the study did not address the worst absolute predicted values. Future research should focus on handling this minor difference appropriately, as it is an important area for further study. The researchers plan to expand their study by using a more detailed approach to forecast container volumes, aiming for even better prediction performance.

REFERENCES

[1] I.Chatterjee and G.S.Cho, "Development of a Machine Learning-Based Framework for Predicting Vessel Size Based on Container Capacity," Applied Science, Vol.12, No.19, pp.1-18, 2022.

[2] I.Chatterjee and G.S.Cho, "Port Container Terminal Quay Crane Allocation, Based on Simulation and Machine Learning Method," Sensors and Materials, Vol.34, No.2 , pp.843-853, 2022.

[3] S.J.Ko and S.J.Kim, "A Study on Pipeline Design Methods for Providing Secure Container Image Registry," Journal of Internet of Things and Convergence, Vol.9, No.3, pp.21-26, 2023.

[4] T.K.Kim, "IoT (Internet of Things)-based Smart Trash Can," Journal of Internet of Things and Convergence, Vol.6, No.1, pp.17-22, 2020.

[5] T.K.Kim, "Self-powered Wireless Bus Information and Disaster Information System based on Internet of Things (IoT)," Journal of Internet of Things and Convergence, Vol.8, No.1, pp.17-22, 2022.

[6] D.H.Kim and K.B.Lee, "Forecasting the Container Volumes of Busan Port using LSTM," Journal of Korea Port Economic Association, Vol.36, No.2, pp.53-62, 2020.

[7] G.D.Yi, "Forecasting the Container throughput of the Busan Port using a Seasonal Multiplicative ARIMA Model," Journal of Korea Port Economic Association, Vol.29, No.3, pp.1-23, 2013.

[8] E.J.Lee, D.H.Kim and H.R.Bae, "Container Volume Prediction using Time-Series Decomposition with a Long Short-Term Memory Models," Applied Sciences, Vol.11, No.19, 8995, 2021.

[9] S.Filom, A.M.Amiri and S.Razavi, "Applications of Machine Learning Methods in Port Operations-A Systematic Literature Review," Transportation Research Part E: Logistics and Transportation Review, Vol.161, 102722, 2022.

[10] D.Tarwidi, S.R.Pudjaprasetya, D.Adytia and M.Apri, "An Optimized XGBoost-based Machine Learning Method for Predicting Wave Run-up on a Sloping Beach," MethodsX, Vol.10, 102119, 2023.

[11] T.Chen and C.Guestrin, "Xgboost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp.785-794, 2016.

[12] G.S.Cho, "Application of Shared Logistics Operation Platform of ICT-based Logistics Companies," Journal of Internet of Things and Convergence, Vol.8, No.5, pp.27-31, 2022.

[13] C.S.Bojer and J.P.Meldgaard, "Kaggle Forecasting Competitions: An Overlooked Learning Opportunity," International Journal of Forecasting, Vol.37, No.2, pp.587-603, 2021.

[14] A.D.Myttenaere, B.Golden, B.L.Grand and F.Rossi, "Mean Absolute Percentage Error for Regression Models," Neurocomputing, Vol.192, pp.38-48, 2016.

[15] N.G.Divergences, "WORLD ECONOMIC OUTLOOK," World Economic Outlook, 2023.

웬티프엉타인(Nguyen Thi Phuong Thanh) [준회원]

- 2021년 2월 : 동명대학교 국제물류학과 (경영학사)
- 2023년 3월 ~ 현재 : 동명대학교 항만물류시스템학과 (공학석사)

〈관심분야〉
물류시스템, 항만물류, 시뮬레이션, 머신 러닝

조 규 성(Gyu Sung Cho) [정회원]

- 1998년 2월 : 동의대학교 산업공학과(공학사)
- 2000년 2월 : 동의대학교 산업공학과(공학석사)
- 2003년 2월 : 동의대학교 산업공학과(공학박사)
- 2012년 3월 ~ 현재 : 동명대학교 항만물류시스템학과 조교수

〈관심분야〉

물류시스템, 항만물류, 시뮬레이션