ICCEPM 2024 https://dx.doi.org/10.6106/ICCEPM.2024.0271
The 10th International Conference on Construction Engineering and Project Management
Jul. 29-Aug.1, 2024, Sapporo

The optimal method for imputing missing data in the preprocessing phase to enhance the performance of a DNN-based construction period prediction model

Haneul LEE[1]*, Yeongchae YUN[2], Youkyung KIM[3], Seokheon YUN[4]

[1*] *Department of Architecture Engineering, Faculty of Engineering, University of Gyeongsang National, South Korea,* E-mail address: gksmf@gnu.ac.kr
[2] *Department of Architecture Engineering, Faculty of Engineering, University of Gyeongsang National, South Korea,* E-mail address: qwasok1029@gnu.ac.kr
[3] *Department of Architecture Engineering, Faculty of Engineering, University of Gyeongsang National, South Korea,* E-mail address: yukyung1229@gnu.ac.kr
[4] *Department of Architecture Engineering, Faculty of Engineering, University of Gyeongsang National, South Korea,* E-mail address: gfyun@gnu.ac.kr

Abstract: The success of construction projects is influenced by various factors, with accurate management and prediction of the construction period playing a crucial role. The construction period is determined through contracts between the client and the contractor, and it is considered a key element in the management of construction projects, alongside cost management. To ensure the successful completion of projects, accurate prediction of the construction period is essential, as it aids in the efficient allocation of time and resources. The main objective of this study is to maximize the performance of construction period prediction models by applying and comparing various methods for handling missing data. Optimizing the model's performance requires accuracy and completeness of data, with the process of outlier removal and missing data imputation potentially having a significant impact on the model's predictive capability. During this process, the effect of changes in the dataset on model performance will be closely examined to identify the most effective method for handling missing data. Outlier removal and missing data imputation are crucial steps in the data preprocessing phase, and they can significantly improve the model's accuracy and reliability. This research aims to apply these data preprocessing methods and analyze their outcomes to find the most effective missing data imputation method for construction period prediction. After the selection process, considering the model's performance and stability, the mode imputation method was identified as the most suitable for predicting the construction period. The findings of this research are expected to contribute not only to improving the accuracy of construction period predictions but also to enhancing the overall efficiency and success rate of construction project management.

Key words: construction period, machine learning, missing data, imputation methods

1. INTRODUCTION

The determination of construction periods in construction projects is made through contracts between the client and the contractor, serving as a fundamental component of project management alongside cost management[1]. Thus, the accurate prediction of construction periods occupies a significant importance in construction projects. Recently, with the advancement of machine learning technology, its application across various sectors within the construction industry has been increasing[2], especially in the area of construction period prediction, where research utilizing machine learning is actively being conducted. For this, a substantial amount of data suitable for machine learning models is essential. However, due to the human involvement in collecting data from construction sites, there is a high probability of data errors such as missing values and outliers caused by human error. For these reasons, the importance of data preprocessing has been widely recognized for its impact on accuracy and prediction performance for a long time[3]. Effective and rational analysis necessitates research into large-scale data preprocessing techniques[4]. Preprocessing the dataset used in improving the performance of construction period prediction models is critical, especially the development of preprocessing techniques through the analysis of information embedded within construction period data.

This study will focus on investigating missing data imputation in the preprocessing phase of the dataset used for enhancing the performance of construction period prediction models. Commonly used methods for missing data imputation, such as mean imputation, median imputation, and mode imputation, will be applied and compared to assess the performance of machine learning models used for construction period prediction. Additionally, the study will observe changes in the dataset size through outlier removal processes and evaluate the potential for analyzing patterns within the data.

2. LITERACTURE REVIEW

Cho B.N., Kim H.S., and Kang L.S. aimed to develop a model capable of estimating work durations by applying neural network theory, to consider various factors that can occur in construction sites comprehensively. By utilizing combinations of different influencing factors, the suitability was reviewed, and the optimal structure of influencing factors was derived. Based on this analysis, a methodology for constructing a model to estimate work durations for each construction process based on neural networks was proposed to overcome the limitations of existing work duration estimation methods. The validation of the model through the comparison of predicted values and actual values showed a mean absolute error rate of 9.8% and a mean coefficient of determination of 0.939, confirming the model's suitability and applicability[5].

Kim S. W. conducted a study on a real-case project of a steel-reinforced concrete building in Korea, ranging from six underground to fifty above-ground floors. The study compiled a dataset by collecting data on construction volume, quantity of resources deployed, planned duration, actual duration, and other factors influencing the construction period for a benchmark floor. Based on the collected data, an analysis of the structural work data from the construction site of a high-rise office building in Korea was performed. The study utilized Deep Neural Networks (DNN) to predict the construction period based on key factors for standard floors in core wall construction. In addition to the factors used as input variables in the DNN model, a case analysis was conducted to explore the impact of other construction period influencing factors, thereby verifying the suitability of AI models like DNN for construction period predictions. The predictions closely matched the actual durations against the planned schedules, with the developed DNN model showing an accurate prediction with only about a 3-day discrepancy in a total construction period of 151 days[6].

Yun H. S. limited the scope of the study to 15 buildings within 10 projects of domestic high-rise curtain wall construction, including actual case process management data based on quantities and resources for main uses such as residential buildings with 53 and 85 floors, a hotel with 101 floors, and office facilities with 72 floors. The research utilized Monte Carlo simulations to input variables for 15 process management data across 10 projects, running 10,000 simulations to predict and compare the distribution of the total project duration for curtain walls, based on the fit of distributions, and the application of triangular and PERT distributions[7].

Rashid, Khandakar M., and Joseph Louis developed a model to predict the final construction contract duration at the planning stage. They created a construction duration prediction model using artificial neural networks (ANN) based on data from 135 Saudi construction projects. The developed ANN model was compared with three linear regression models and models presented in the literature. The analysis revealed that the ANN model had the lowest error rate, with a Mean Absolute Percentage Error (MAPE) of 12.22%, proving to be the most accurate in predicting actual contract durations. This accuracy is expected to assist in making appropriate decisions about projects in the pre-bidding phase[8].

Mahmoodzadeh, Arsalan, Hamid Reza Nejati, and Mokhtar Mohammadi aim to compare models for predicting the duration and cost of tunnel construction projects and to present the model with the highest performance. For this purpose, machine learning regression models incorporating Linear Regression (LR), Gaussian Process Regression (GPR), Support Vector Regression (SVR), and Decision Tree (DT) were utilized. These models predicted the duration and cost of tunnel construction based on 16 input variables. The prediction performance was ranked from highest to lowest as follows: LR, GPR, SVR, and DT. Sensitivity analysis indicated that the drilling machine system and groundwater were the factors most significantly affecting the duration and cost of tunnel construction[9].

Sanni-Anibire, Muizz O., Rosli Mohamad Zin, and Sunday Olusanya Olatunji aim to develop a model for predicting the construction duration of high-rise building projects. For model development, Multi-

Linear Regression Analysis (MLRA), k-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and ensemble techniques were utilized, resulting in a total of 12 models. After comparing the performance of these 12 models, the models using ANN and ensemble techniques were identified as having the best prediction performance. These models showed a correlation coefficient of 0.69, a Root Mean Square Error (RMSE) of 301.72, and an average error rate of 18%, indicating a considerable accuracy in predicting the construction duration[10].

3. ANALYSIS FOR OPTIMAL IMPUTATION METHOD

3.1. Data Collection

There are difficulties associated with collecting performance data in construction projects, and there is a propensity for outliers and missing values due to human error. If artificial intelligence training is conducted based on such incomplete data, there is a risk of incorrect interpretations of the inherent data patterns. Therefore, to accurately grasp the inherent data patterns, it is crucial to collect as much performance data as possible. In this study, a total of 2318 cases of construction duration performance data were used, compiled based on information collected from various projects and sources. The data collection did not limit the scope to the purpose or size of the buildings, aiming to gather as much information as possible.

An analysis of the collected data revealed the number of missing values for each item: 'Total Area' had 6 missing cases out of 2318, 'Building Area' had 105 missing cases out of 2318, 'Basement' had 68 missing cases out of 2318, and 'Groundlevel' had 73 missing cases out of 2318. This analysis indicates that certain items have a relatively high number of missing values that need to be addressed.

3.2. Construction Period Prediction Model

The construction duration prediction model is designed to identify the correlation between four key influencing factors—'Total Area', 'Building Area', 'Basement', 'Groundlevel'—and the construction duration, aiming to predict the construction time based on these data patterns. These factors are essential variables that have a proportional relationship with the construction period.

The development environment for the model is 'Visual Studio Code', and it was developed using 'Python'. The machine learning model developed in this research is focused on predicting construction duration, with the scope limited to the data preprocessing step of replacing missing values. The methods used for missing value replacement include the most commonly used techniques: mean replacement, median replacement, and mode replacement. The model employs the 'ELU' activation function, with a node progression of 100->64->32->1, and the learning iterations set to 300 for all three cases, only varying the method of missing value replacement.

Table 1. Base Model Configuration

Case.	Activation Function	Epoch	Node	Missing Data Imputation Method
A	ELU	300	100, 64, 32, 1	Mean Imputation
B	ELU	300	100, 64, 32, 1	Median Imputation
C	ELU	300	100, 64, 32, 1	Mode Imputation

4. CASE ANALYSIS

Due to the prevalence of outliers and missing values caused by human error in the collected data, a process for replacing missing values is necessary. In this study, mean replacement, median replacement, and mode replacement methods were applied as methods for replacing missing values. The study presents an appropriate method for missing value replacement through the analysis of graphs showing the number of outliers identified and removed after replacement, and the performance of the learning model.

Table 2 Change in Dataset Size Due to Data Preprocessing Steps

Case.	Missing Data Imputation Method	Before Data Preprocessing	After Data Preprocessing
A	Mean Imputation	2318	2212
B	Median Imputation	2318	2223
C	Mode Imputation	2318	2218

Table 2 demonstrates the change in the number of data after detecting and removing outliers with each replacement method. In 'CASE A', the total number of data decreased from 2318 to 2212, reducing by 106. In 'CASE B', it decreased by 95 to 2223, and in 'CASE C', it decreased by 100 to 2218. Therefore, among the three cases, 'CASE B' showed the least reduction in data after outlier removal, followed by 'CASE C' and 'CASE A'.

Figure 1 presents the learning outcome for 'CASE A', which involves the application of the mean replacement method to the machine learning dataset for construction duration prediction. The analysis indicates that as the Epoch increases, both the validation and training values tend to converge to 0. However, significant noise is observed in both values, suggesting that the learning performance is unstable.

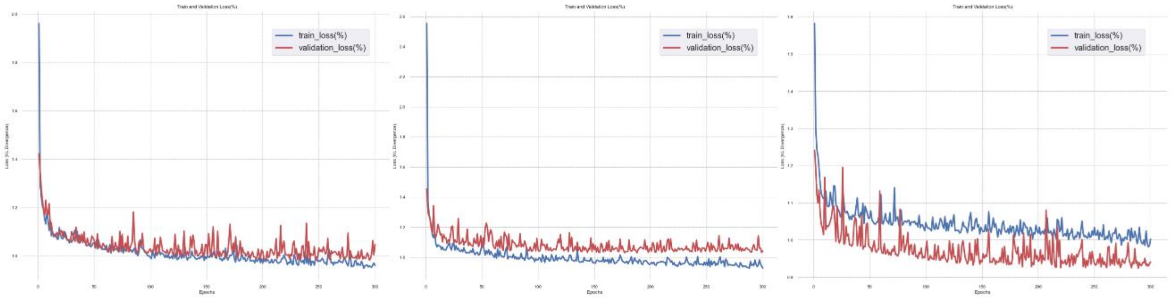

Figure 1 Training Results for ‘CASE A’

Figure 2 shows the learning results for 'CASE B', which applies the median replacement method to the machine learning dataset for predicting construction duration. According to the analysis, although the Epoch increases, both the validation and training values decrease only initially and then maintain a constant level towards the latter half, not showing further convergence. Compared to 'CASE A', there is less noise, suggesting that the learning performance is relatively stable.

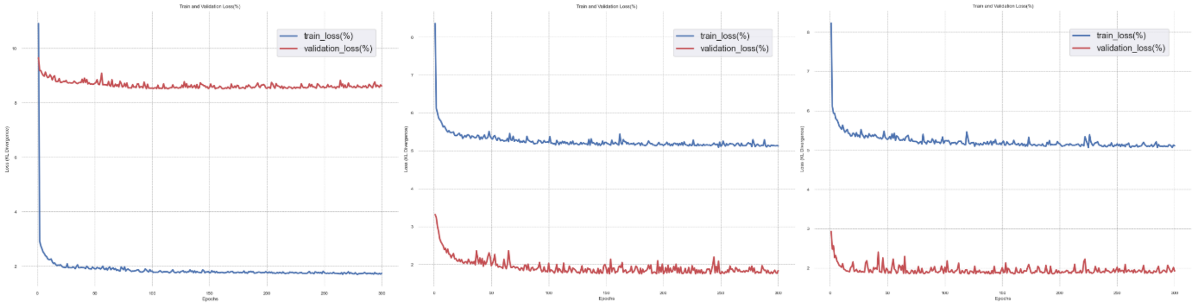

Figure 2 Training Results for ‘CASE B’

Figure 3 presents the learning outcomes for 'CASE C', which involves applying the mode replacement method to the machine learning dataset for construction duration prediction. The analysis indicates that at Epoch 50, both the validation and training values show a sharp decrease towards 0, and as the Epoch progresses, both values gradually converge to 0. Unlike 'CASE A' and 'CASE B', 'CASE C' exhibits relatively less noise, and the gap between the training and validation values narrows, showing a converging trend. This suggests that the learning performance is stable and superior in 'CASE C'.

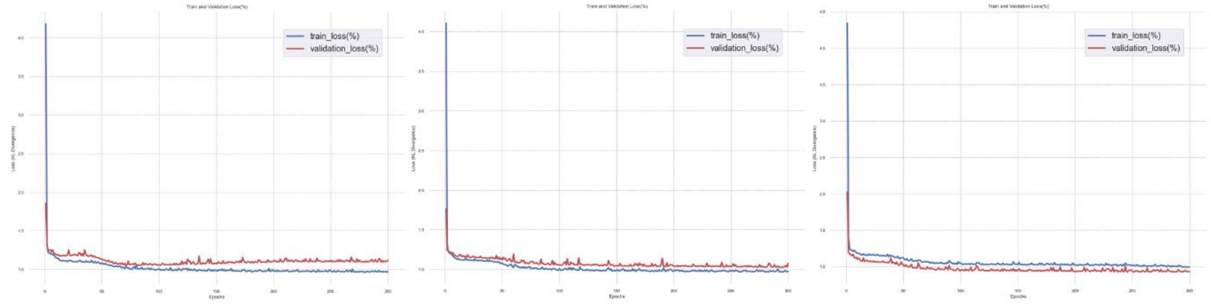

Figure 3 Training Results for 'CASE C'

5. CONCLUSTION

The primary objective of this study is to apply various methods to replace missing values in the performance dataset used for training the construction duration prediction model and to select the most suitable method among them. For this purpose, mean replacement, median replacement, and mode replacement methods were applied, followed by outlier removal to observe changes in the dataset and compare performance to determine the most appropriate missing value replacement method. In order of the least number of data identified and removed as outliers after replacement, the sequence is 'CASE B', 'CASE C', 'CASE A'. Analysis of the learning outcome graphs in terms of stability and performance showed that 'CASE C' exhibited the most stable and excellent learning performance.

Therefore, it is concluded that the mode replacement method applied in 'CASE C' is deemed the most appropriate for replacing missing values in the construction duration performance data.

ACKNOWLEGEMENTS

This work is supported in 2024 by the Korea Agency for Infrastructure Technology Advancement(KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant RS-2021-KA163269).

REFERENCES

[1] Kim, Eunju, et al. "Analysis of Different Perception on the importance of Construction Schedule Delay Factors by Participants in Apartment Construction Projects." Journal of the Regional Association of Architectural Institute of Korea 18.3 (2016): 165-172.

[2] Kim, Gyu-Tae, and Seok-Heon Yun. "Impact Analysis of Drop out Method on Machine Learning-based Construction Cost Estimation." The Journal of Next-generation Convergence Technology Association 6.4 (2022): 641-647.

[3] Kim, S. H., Suh, Y. K., Tak, B. C. (2020). A Recommendation Scheme for an Optimal Pre-processing Permutation Towards High-Quality Big Data Analytic. Journal of KIISE, 47(3), 319-327.

[4] Jun, H. K., Hyun, G. S., Lim, K.B., Lee, W. H., & Kim, H. J. (2014). Big Data Preprocessing for Predicting Box Office Succes. KIISE Transactions on Computing Practices, 20(12), 615-622.

[5] Cho, Bitna, et al. "Development of Estimation Model of Construction Activity Duration Using Neural Network Theory." Journal of the Korea Academia-Industrial cooperation Society 16.5 (2015): 3477-3483.

[6] Kim, Seung-Woo. "A Study to Predict the Duration of Building Structural Works By Introducing Deep Neural Network." Master's Thesis, Hanyang University's Graduate School of Engineering in Seoul, 2021.

[7] Yun, Hye-Sun. "A Study to Predict the Duration of Curtain Wall Works in High-rise Building by Introducing Monte Carlo Simulation & Machine Learning." Master's Thesis, Hanyang University in Seoul, 2021.

[8] Khandakar M, R., & Louis, J. (2019). Times-series data augmentation and deep learning for construction equipment activity recognition. Advanced Engineering Informatics, 42, 100944.

[9] Mahmoodzadeh, A., Nejati, H. R., & Mohammadi, M. (2022). Optimized machine learning modelling for predicting the construction cost and duration of tunnelling projects. Automation in Construction, 139, 104305.

[10] Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2021). Developing a machine learning model to predict the construction duration of tall building projects. Journal of Construction Engineering, 4(1), 022-036.