International Journal of Naval Architecture and Ocean Engineering 12 (2020) 428–439

Contents lists available at ScienceDirect

International Journal of Naval Architecture and Ocean Engineering

journal homepage: http://www.journals.elsevier.com/international-journal-of-naval-architecture-and-ocean-engineering/

Machine Learning Methodology for Management of Shipbuilding Master Data

Ju Hyeon Jeong [a], Jong Hun Woo [b, *], JungGoo Park [c]

[a] Korea Maritime and Ocean University, South Korea
[b] Seoul National University, South Korea
[c] Ship & Offshore Research Institute, Samsung Heavy Industries, South Korea

ARTICLE INFO

Article history:
Received 19 June 2019
Received in revised form
2 February 2020
Accepted 3 March 2020

Keywords:
Big data
Statistical analysis
Machine learning
Shipbuilding
Production management
Master data

ABSTRACT

The continuous development of information and communication technologies has resulted in an exponential increase in data. Consequently, technologies related to data analysis are growing in importance. The shipbuilding industry has high production uncertainty and variability, which has created an urgent need for data analysis techniques, such as machine learning. In particular, the industry cannot effectively respond to changes in the production-related standard time information systems, such as the basic cycle time and lead time. Improvement measures are necessary to enable the industry to respond swiftly to changes in the production environment. In this study, the lead times for fabrication, assembly of ship block, spool fabrication and painting were predicted using machine learning technology to propose a new management method for the process lead time using a master data system for the time element in the production data. Data preprocessing was performed in various ways using R and Python, which are open source programming languages, and process variables were selected considering their relationships with the lead time through correlation analysis and analysis of variables. Various machine learning, deep learning, and ensemble learning algorithms were applied to create the lead time prediction models. In addition, the applicability of the proposed machine learning methodology to standard work hour prediction was verified by evaluating the prediction models using the evaluation criteria, such as the Mean Absolute Percentage Error (MAPE) and Root Mean Squared Logarithmic Error (RMSLE).

1. Introduction

1.1. Background

The widespread growth of the information society has resulted in an exponential increase in data, which has led to the rapid development of big data technology for systematic collection, storage, and analysis of large volumes of data that are difficult to collect, store, and analyze using the existing methods or tools. In the era of the fourth industrial revolution, there have been increasing attempts to create new added values through big data technology, mainly in industries.

Big data has been applied to the shipbuilding industry in various ways. Studies on predicting the types of ships through the analysis of sales data from the international market (Lee, 2017) and studies on predicting quantity through the analysis of BOM and design information can be referred to as representative studies (Oh et al., 2018). In the production management area, studies on standard work hours and standard lead time, which are the concepts addressed in this study, have been conducted (Ham, 2016).

In the existing method for standard work hour management, the quantity is calculated from the product information and the work hours and lead time are analyzed by adding the existing standard unit work hours to the calculated quantity. It can be said that the method is based on causal relation. This method, however, requires further research on quantity calculation or standard unit work hours owing to the complexity of products and the existence of various variables in the field work despite the research being conducted over an extended period.

The big data analysis methodology, on the other hand, is an analysis technique based on statistics, which analyzes only the correlations between the data expressed in numbers or letters

* Corresponding author.
E-mail addresses: j.woo@snu.ac.kr, jonghun.digitalfactory@gmail.com (J.H. Woo).

Peer review under responsibility of Society of Naval Architects of Korea.

https://doi.org/10.1016/j.ijnaoe.2020.03.005
2092-6782/

regardless of the products or process technologies (Lee et al., 2014a,b). Therefore, it is expected that a standard work hour prediction model for the target work can be derived through machine learning of the relations between relevant variables and actual (or planned) data if the big data analysis methodology is applied to the standard work hours.

1.2. Current problems

The production time of most general manufacturing industries (mechanical, aviation, electronics, etc.), which are mainly operated by mass production systems, has a high standardization rate. In the case of workers, standardized motion analysis methods such as MTM (Methods-Time Measurement) and MODAPTS (MODular Arrangement of Predetermined Time Standards) are used. In the case of facilities, standard operation sheet analysis is used to standardize the working time and reflect them in the process design. In the general manufacturing industry, the standardization of such working time is possible because the same (or similar) product is produced repeatedly. Except, the same applies.

However, the shipbuilding is very difficult to standardize the working time because the specifications of the intermediate product (hull block, design, member, etc.) constituting the final product as well as the final product (ship) are all different.

Once the ship type is determined to be built at the shipyard, the production plan is usually based on the previous vessel's plan data rather than the standard of estimate information on production time. In principle, the amount of work should be calculated from the product information of the ordered vessel and the working time should be calculated in consideration of the standard of estimate. However, due to the wide variety and quantity of vessels, this principle is not well followed, so the accuracy rate of the production plan is low compared to the other manufacturing sectors.

In this study, we propose a model that can predict more accurate working time through supervised learning on past production record as a solution for the problem of production planning.

2. Objectives

In this study, a model for predicting the standard time data of a shipyard was created using the big data analysis methodology. The scenarios related to shipbuilding production were defined by collecting the performance data of shipyards on fabrication, assembly, and procurement, and were analyzed using the learning algorithms in R and Python.

Machine learning, deep learning, and ensemble learning algorithms were applied to the fabrication, assembly, and procurement data related to shipbuilding in this study. Lead time prediction models were created based on analytical cases by applying various learning algorithms. The reason for applying various learning algorithms is to investigate whether a specific algorithm is good for all process data or specific algorithms guarantee high prediction accuracy for a certain process type, because the prediction accuracy of the learning model can vary depending on the characteristics of the production data. . The prediction models were evaluated by comparing the lead times obtained from the models with the actual lead time.

3. Related studies

Jo and Kang (2016) classified the big data applicable to the manufacturing industry under product development, manufacturing process, sales and marketing, and warranty service, and then presented application examples for each field. In the text mining case for improving the efficiency of the automotive parts design process in the product development area, they showed that the time and effort required to identify major problems could be reduced by analyzing the information included in the design verification test.

Jung and Sim (2014) attempted to reduce the welding cost by analyzing the work patterns of welders based on the welding data. To this end, they examined the patterns of power consumption and the wire length consumed by applying the algorithm in R and regression analysis to a large number of welding work pattern variables.

Ham (2016) and Ham et al. (2016) conducted a research on improving the level of procurement management by predicting the lead time for spools, which are outfittings that cause significant delays in post tasks, from the manufacturing process to the installation process. In the study, the lead time was defined by dividing the supply chain of the piping process into six processes, and multiple linear regression analysis and partial least squares regression analysis were conducted. However, the error rate of the lead time for each process was found to be large because of insufficient data preprocessing.

Hur et al. (2015) analyzed the effort data of the ship design and production processes to predict the effort in shipyards. Prediction models were created by defining variables related to effort and using multiple linear regression analysis as well as decision tree. However, the method suffered from limitations related to the collection of long-term data considering that shipbuilding in shipyards takes one to two years on average, and the consideration of external factors other than the workspace was inadequate.

Lee et al. (2014a,b) applied the text mining method to predict various kinds of defects that could occur during the construction of marine structures. They extracted significant knowledge helpful for the manufacturing process by conducting defect trend analysis and related defect analysis through the analysis of text log data and visualization of results.

National Information Society Agency (NIA) (2016) pushed forward a project for developing a big data cloud service for analyzing the shipyard manufacturing process as part of the big data pilot project. The project attempted to improve the work efficiency by analyzing process delays and loads. The process status and delay factors were identified by examining the manufacturing process big data based on process mining.

Kim et al. (2016) proposed a big data platform based on Hadoop, a big data distributed processing technology. They studied the applicability of big data to marine structure development by applying the proposed platform to the estimation of the weight of the offshore plant superstructure. However, the data for analysis were not sufficient owing to security concerns of shipyards, and the testing of various analysis algorithms was inadequate because only simple linear regression analysis was applied.

Currently, big data analysis research continues to increase in the shipbuilding and marine industries, but no research case has been found in the field of lead time related to shipyard production. There are also cases where technologies related to big data have been acquired through various studies, but there are no applicable objects (actual shipyard data). Therefore, in this study, we intend to predict the lead time of production by applying machine learning methodology to actual shipyard data.

4. Algorithms used for analysis

In this study, machine learning, deep learning, and ensemble learning were used as analytic algorithms. Deep learning and ensemble learning are included in one of the machine learning methodologies, but we will use them separately for convenience to show the process of expanding the analytic algorithm.

4.1. Machine learning algorithm

Machine learning, an application of artificial intelligence, can be referred to as a technology for constructing ideal learning models using various probabilities, combinatorics, mathematical optimization techniques, statistics, and algorithms (Lee et al., 2014a,b). As the purpose of machine learning is to create models using data, selecting appropriate input data as well as selecting an algorithm suitable for a problem is important. Machine learning algorithms are selected based on training data and divided into supervised learning, for which labels are included in the training data, and unsupervised learning, for which no label is included in the training data (Fig. 1).

The purpose of this study is to develop the prediction models for improving the lead times, which are the master data. Therefore, the process data of shipyards were defined as input values and the lead times to be predicted were defined as output values. Thus, the machine learning algorithms of this study were limited to supervised learning algorithms. Among such supervised learning algorithms, multiple regression analysis, single layer perceptron, and decision tree, which are known to be suitable for numerical prediction, were used.

4.1.1. Multiple regression analysis

The most basic algorithm among the numerical prediction algorithms is regression analysis. Regression analysis, which is a statistical analysis method to identify the relationships between variables, is used to predict the values of dependent variables according to the values of independent variables. In this study, multiple regression analysis is used to identify the relationship between one dependent variable and several independent variables.

4.1.2. Single layer perceptron

Single layer perceptron is and early artificial neural network, which consists of sending values and outputting values. It receives multiple signals and outputs one signals, which seems similar to the neuron sending out an electrical signal to transmit information. The single layer perceptron receives the signal as an input and forwards the information of 1 or 0 and gives unique weight to each of the multiple input signals.

4.1.3. Decision tree

Decision tree, an algorithm based on inductive inference, is one of the most commonly used supervised learning models in the field. Because the analytical process is expressed using a tree structure, researchers can easily understand and explain the analysis process. Although the decision tree is a typical classification model, it is classified as a classification tree if the target variable is categorical and a regression tree if the target variable is continuous. In this study, the analysis is performed as a regression tree because the target variable is lead time which is a continuous variable.

4.2. Deep learning algorithm

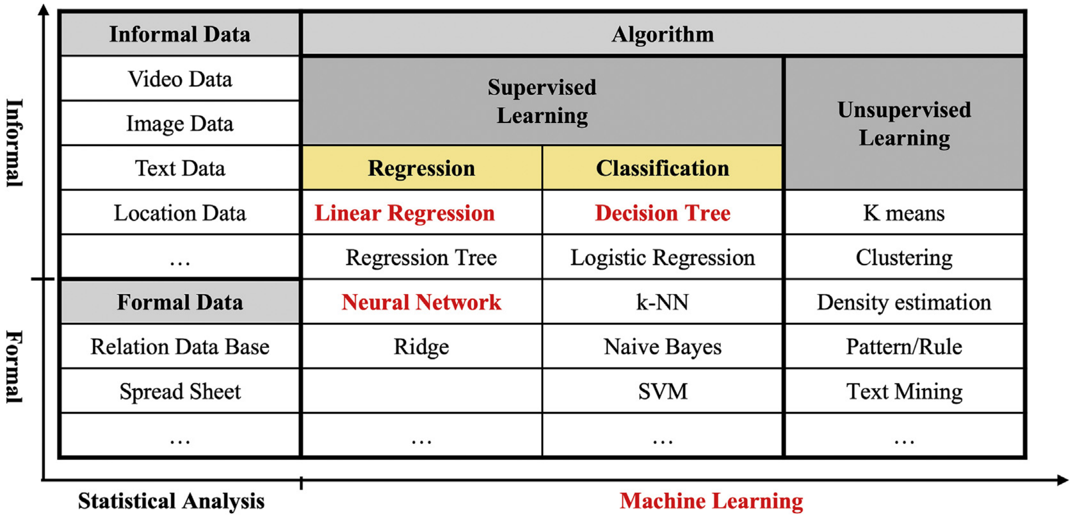
In the existing machine learning process, people frequently designate definitions in advance or human interaction is involved when the computer extracts features from the training data, which results in many errors. Moreover, the approach of using multiple neural layers has not been utilized owing to problems, such as nonlinearity, limitations on the number of weights due to the number of layers, and overfitting. Despite such problems, deep learning technology is being widely used in artificial intelligence because improved computational performance of computers and the development of algorithms have demonstrated the usefulness of multi-layered neural networks (Kim et al., 2016).

4.2.1. Multi-layer perceptron

Multi-layer perceptron was proposed as a way to overcome the limitations of the single layer perceptron that non-linearly separated data were not available for learning. It is possible to learn about data that is non-linearly separated by having one or more hidden layers between the input and output layers. The input layer plays the role of inputting the prediction variables. The hidden layer receives the input values from input nodes, calculates the weights, applies the values to the activation function, and delivers the results to the output layer. In this study, multi-layer perceptron was used for analysis to ensure better performance than single layer perceptron (Fig. 2).

4.3. Ensemble learning algorithm

Ensemble learning is a method of learning a new hypothesis by learning several single classifiers and combining their predictions. The purpose of ensemble learning is to obtain a predicted value



Informal Data	Algorithm		
Video Data	Supervised Learning		Unsupervised Learning
Image Data			
Text Data	Regression	Classification	
Location Data	Linear Regression	Decision Tree	K means
…	Regression Tree	Logistic Regression	Clustering
Formal Data	Neural Network	k-NN	Density estimation
Relation Data Base	Ridge	Naive Bayes	Pattern/Rule
Spread Sheet		SVM	Text Mining
…	…	…	…

Fig. 1. Algorithm classification of machine learning.

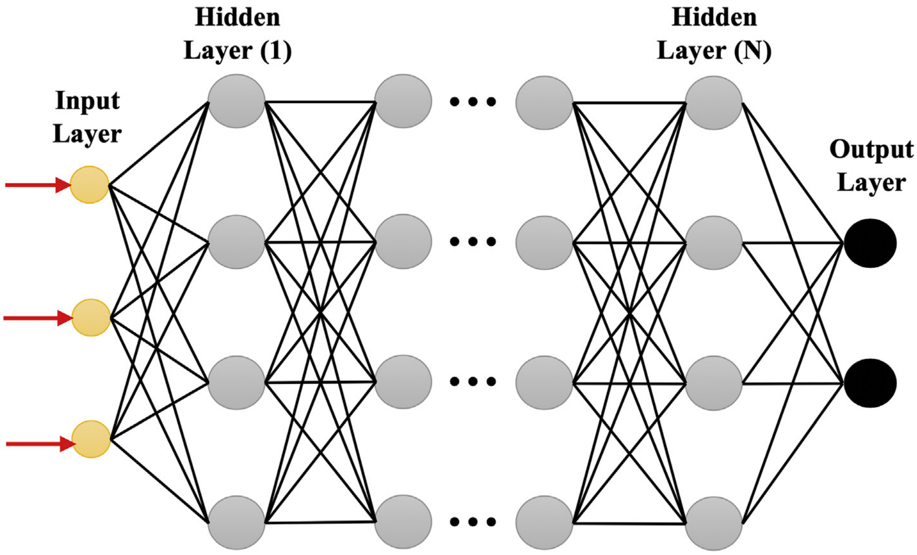


Fig. 2. Concept of multi-layer perceptron.

with higher reliability than that obtained with a single classifier by combining the results of various classifiers (Lee and Yang, 2008). To obtain excellent results from ensemble learning, it is necessary to represent the classifiers using various algorithms or to randomly divide the training dataset and train each classifier differently even though the same algorithm is used.

Ensemble learning is a classified as bagging, boosting and stacking. Bagging uses one learning algorithm. And this is a method of allowing duplication in training data sets to learn different learning models and make decisions by aggregating each learning result. Boosting is the same as bagging, but it is characterized by a higher weight on the wrong answer and a lower weight on the correct answer, thereby focusing more on the wrong answer. Stacking is an algorithm that combines different learning models. As a result, the advantages of each algorithm are taken and the weaknesses of each algorithm are compensated. In this study, random forest corresponding to bagging is used for analysis.

4.3.1. Random forest algorithm

In this study, the random forest algorithm was used for ensemble learning. Random forest is an algorithm obtained by improving the decision trees of machine learning. It creates a model by combining multiple decision trees. The concept of the Random forest algorithm is illustrated in Fig. 3. Random forest selects variables by using the bagging method, in which sampling is performed by allowing duplication of the training dataset, for the construction of each decision tree. Each unit model is significantly different because they have different independent variables, and it is possible to obtain a predicted value with higher reliability than that possible with a single model by combining the prediction results of each model. Unlike the existing ensemble model, random forest is stable because it maximizes the benefits of ensemble learning and improves the prediction and classification accuracy by applying randomness to variables as well as the observed values. Random forest constructs the final ensemble learning model, $C^*(x)$, by combining B decision tree models. The method of constructing the final learning model is different depending on the analysis target.

In the regression model, the method for constructing the final learning model involves the calculation of the average of the values predicted by each decision tree, as follows.

$$C^*(x) = \sum_{b=1}^{B} C_b(x)/B$$

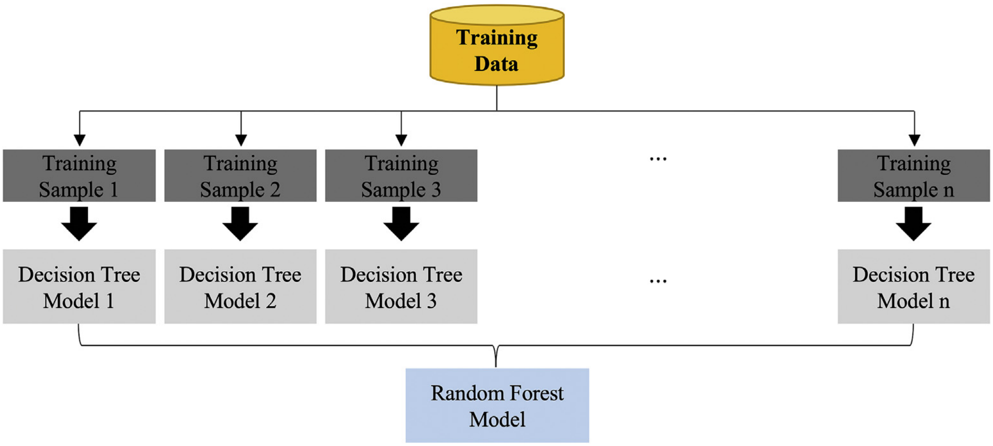


Fig. 3. Concept of random forest algorithm.

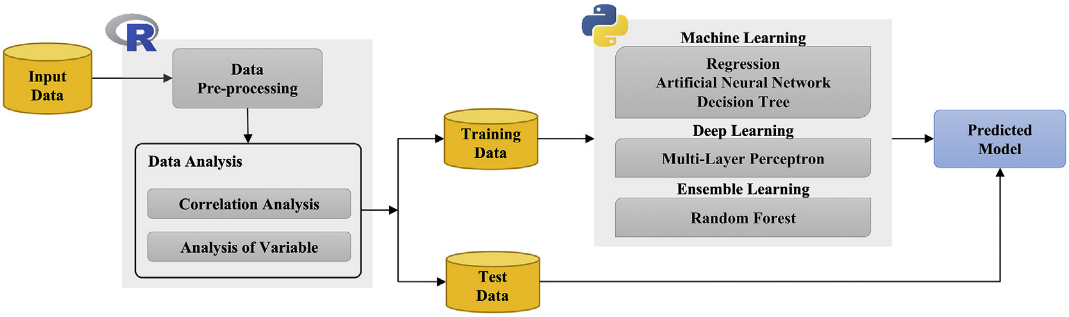


Fig. 4. Process of data analysis.

In the classification model, the method for constructing the final learning model involves identifying the most selected class by voting.

$$C^{*}(x) = argmax_{y} \sum_{b=1}^{B} I[C_b(x) = y]$$

5. Data analysis process

The data analysis process comprises four steps: data collection, data processing, model construction, and model performance validation (Fig. 4). In the data collection step, the data pertaining to a shipyard for which the production lead time is to be predicted are collected for each process and various data processing techniques for constructing a model are applied. In the next step, prediction models are created by classifying the data required for creating the prediction models and by applying learning algorithms. Finally, the performance of each prediction model is validated using the criteria for evaluation of the result data.

5.1. Data collection

In the data collection step, data are collected according to the analysis purpose. The collected shipyard data are classified based on three process types, namely, cutting process, erection process, and spool procurement process, as shown in Table 1. In the case of the spool procurement process, the lead times for spool fabrication and painting, in which the spool process variables have significance, were targeted in this study even though there were various processes between spool fabrication and installation.

Cutting process performance data have been stored for about six years. This data has relatively fewer variables managed together. The collected data is 63,989rows and consists of 7 independent variables (3 continuous variables and 4 categorical variables) and a dependent variable corresponding to lead time. Erection process performance data have been stored for about three years. Continuous variables in the block were extracted by mapping work target and block information. The collected data is 22,758rows and consists of 15 independent variables (8 continuous variables and 7 categorical variables) and a dependent variable corresponding to lead time. Finally, spool procurement performance data are about spools installed in one vessel. The supply chain of the spools is constructed through the process from making to installation. And lead time is controlled for each process, but in this study, the lead time of the making and painting process is predicted. The collected data is 32,039rows and consists of 19 independent variables (6 continuous variables and 13 categorical variables) and a dependent variable corresponding to lead time.

Table 1
Data collection.

Analytical data	Prediction target
Cutting process performance data	Lead time for cutting
Erection process performance data	Lead time for erection process
Spool procurement performance data	Lead times for spool fabrication and painting

5.2. Data processing

In the data processing step, the collected data are defined, searched, modified, and preprocessed. Data analysis involves the tasks of selecting variables to be applied to the algorithms and processing erroneous data.

First, independent and dependent variables are defined for the algorithm. As different analysis techniques can be applied depending on the type and number of variables, independent variables having a significant relationship with the lead time, which is a dependent variable, are selected from among the defined process variables by conducting correlation analysis between continuous variables and the analysis of variables between categorical variables. Next, errors in the analysis results must be prevented in advance by examining each variable and removing outliers or missing values. Missing values in the data corresponding to the selected variables are removed whereas the interquartile range (IQR) rule and Cook's distance are used to remove the outliers.

5.2.1. IQR rule

The IQR rule can be explained by visualizing the distribution of data using a box plot, as shown in Fig. 5. The box plot summarizes the degree of variation of data using the maximum value, upper quartile, median value, lower quartile, and minimum value. IQR is the value obtained by subtracting the lower quartile from the upper quartile. The IQR rule determines the values that exceed the range between the lower quartile $-$ IQR $\times$ 1.5 and the upper quartile $+$ IQR $\times$ 1.5 as outliers. In this study, outliers were removed by applying the IQR rule to the independent and dependent variables, which were continuous variables.

5.2.2. Cook's distance

In regression analysis, values with large leverages and residuals are referred to as outliers. Cook's distance is a criterion for viewing the leverage and residual at the same time. The leverage is a value that represents the influence of the actual result value y on the predicted value $\hat{y}$. For the influence matrix H, $\hat{y} = Hy$ holds and the leverage is mathematically defined as the diagonal component h_{ii} of the influence matrix H. The residual represents the difference between the value estimated from the regression equation of the

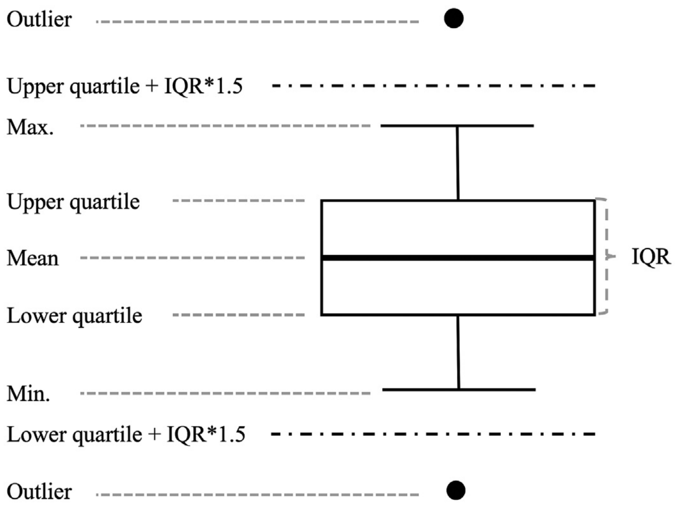


Fig. 5. Visualized IQR rule.

sample and the actual value. Scaled standardized residuals r_i must be used to have the same standard deviation.

Cook's distance is expressed as follows.

$$D_i = \frac{{r_i}^2}{RSS}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

Below is the criterion for determining Cook's distance as an outlier. N is the number of data and K is the sum of leverages.

$$D_i > \frac{4}{N-K-1}$$

In this study, outliers were removed by applying Cook's distance to the lead time, which is a dependent variable.

5.3. Model construction

In this step, the training data for constructing the prediction models and the evaluation data for evaluating the prediction models are classified and the learning algorithms are applied. In general, the learning and evaluation data are classified at a ratio of approximately 7:3. The process data that underwent the data processing step are classified and a learning algorithm is applied to the training data. In addition, the prediction results of the finally created learning model are compared with the evaluation data.

5.4. Model performance validation

In the model performance validation step, the performance of each prediction model is validated using the evaluation criteria. The following evaluation criteria were used to calculate the error between the predicted value and the performance value of the actual data as well as determine the accuracy of the predicted value quantitatively (Table 2.

The mean absolute error (MAE) represents the average of the absolute errors between the predicted values and the actual values. The mean absolute percentage error (MAPE) converts the difference between the predicted value and the actual value into a percentage. The root mean square error (RMSE) is the square root of the average of the squared residuals, and it is usually expressed as precision. Finally, the root mean squared logarithmic error (RMSLE) is the log value of the average of the residuals. Prediction errors may occur in areas with large outliers as well as in areas with small outliers. RMSLE calculates an error by assigning penalties to items with underestimated outliers rather than to items with over-estimated outliers.

Table 2
Evaluation criteria.

Evaluation Criteria	Equation
MAE	$\lvert y_i - \widehat{y}_i\rvert$
MAPE	$\frac{100}{n}\sum_{i=1}^{n}\lvert(y_i - \widehat{y}_i)/y_i\rvert$
RMSE	$\sqrt{[\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2]/n}$
RMSLE	$\sqrt{\sum_{i=1}^{n}(\log(\widehat{y}_i+1) - \log(y_i+1))^2/n}$

Table 3
Results of machine learning (1).

Case	Regression Analysis		Artificial Neural Network		Decision tree	
	1	2	1	2	1	2
MAE	13.47	6.29	12.73	6.16	11.23	5.37
MAPE	167.6%	102.6%	156.8%	101.4%	128.2%	89.3%
RMSE	22.06	8.16	21.39	8.01	19.73	7.41
RMSLE	88.6%	65.9%	84.5%	64.5%	75.3%	60.7%

Table 4
Results of machine learning (2).

Case	Regression Analysis		Artificial Neural Network		Decision tree	
	1	2	1	2	1	2
MAE	13.07	9.96	13.98	8.98	10.19	7.76
MAPE	352.0%	306.1%	391.5%	225.5%	198.0%	182.0%
RMSE	19.62	12.26	20.28	11.76	17.37	10.83
RMSLE	116.2%	107.0%	123.1%	93.6%	88.0%	82.3%

Table 5
Results of machine learning (3).

Case	Regression Analysis		Artificial Neural Network		Decision tree	
	1	2	1	2	1	2
MAE	10.31	6.76	10.97	6.95	9.67	6.26
MAPE	36.1%	29.9%	40.7%	30.8%	32.5%	27.2%
RMSE	22.54	8.55	23.44	8.90	22.45	8.13
RMSLE	39.7%	32.4%	43.1%	33.4%	36.4%	30.4%

Since the four criteria used in this study are relative evaluation criteria for comparing one value with predicted values by various algorithms, it is difficult to present specific criteria that are commonly applied to all data. Therefore, the learning model with the smallest difference between actual lead time and predicted lead time was considered to have relatively good predictive performance. In this study, it was determined that MAPE was the most intuitive way to identify the effect of error on the data. Therefore, MAPE was considered first in the model performance validation step.

6. Data analysis for predicting the shipbuilding production lead time

In this study, production process data were collected from shipyards to apply to various learning algorithms, and the data

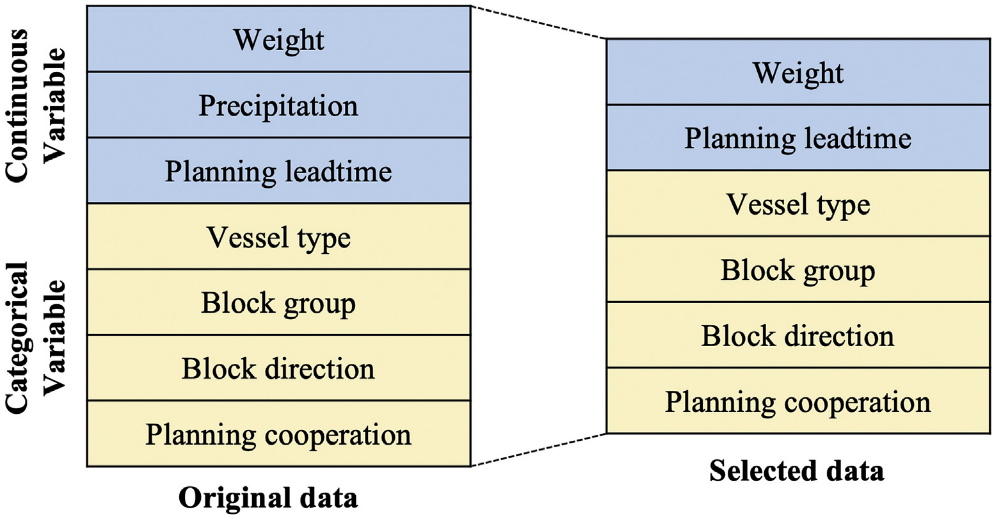


Fig. 6. Independent variables (1).

were analyzed according to the requirements of shipbuilding production management to define various scenarios. Three cases are presented in this study: the lead time for cutting, the lead time for block erection, and the lead time for spool procurement in shipyards.

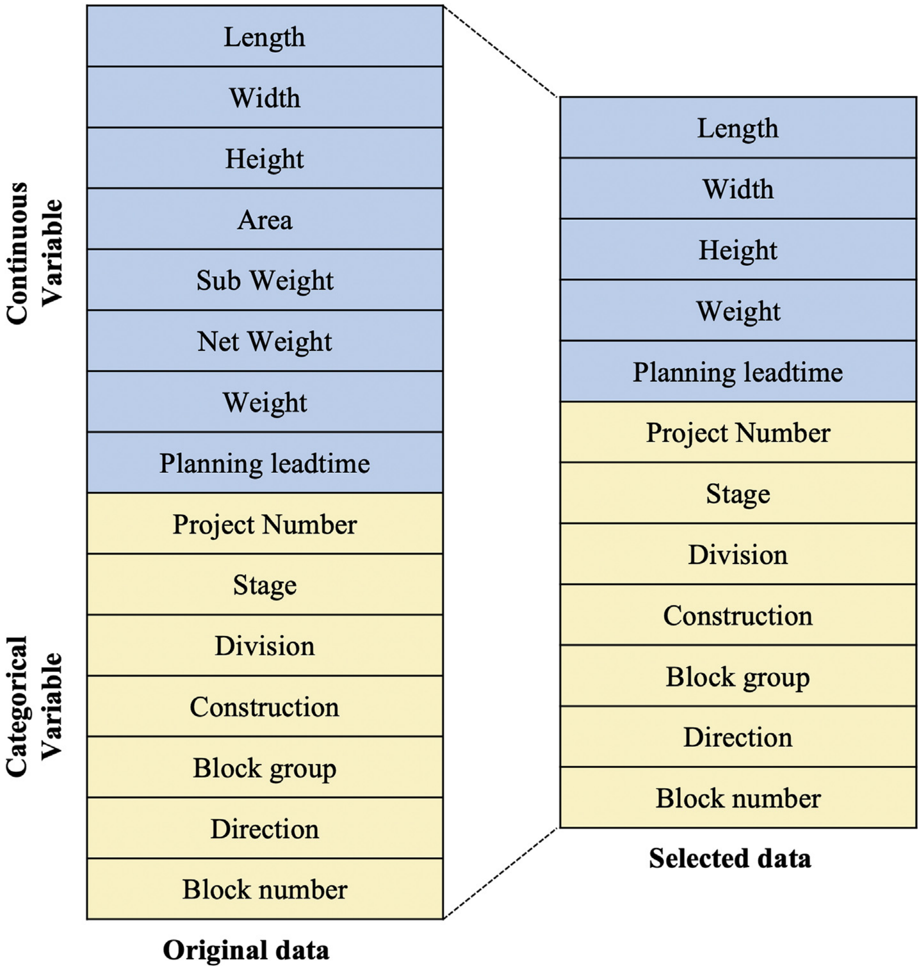


Fig. 7. Independent variables (2).

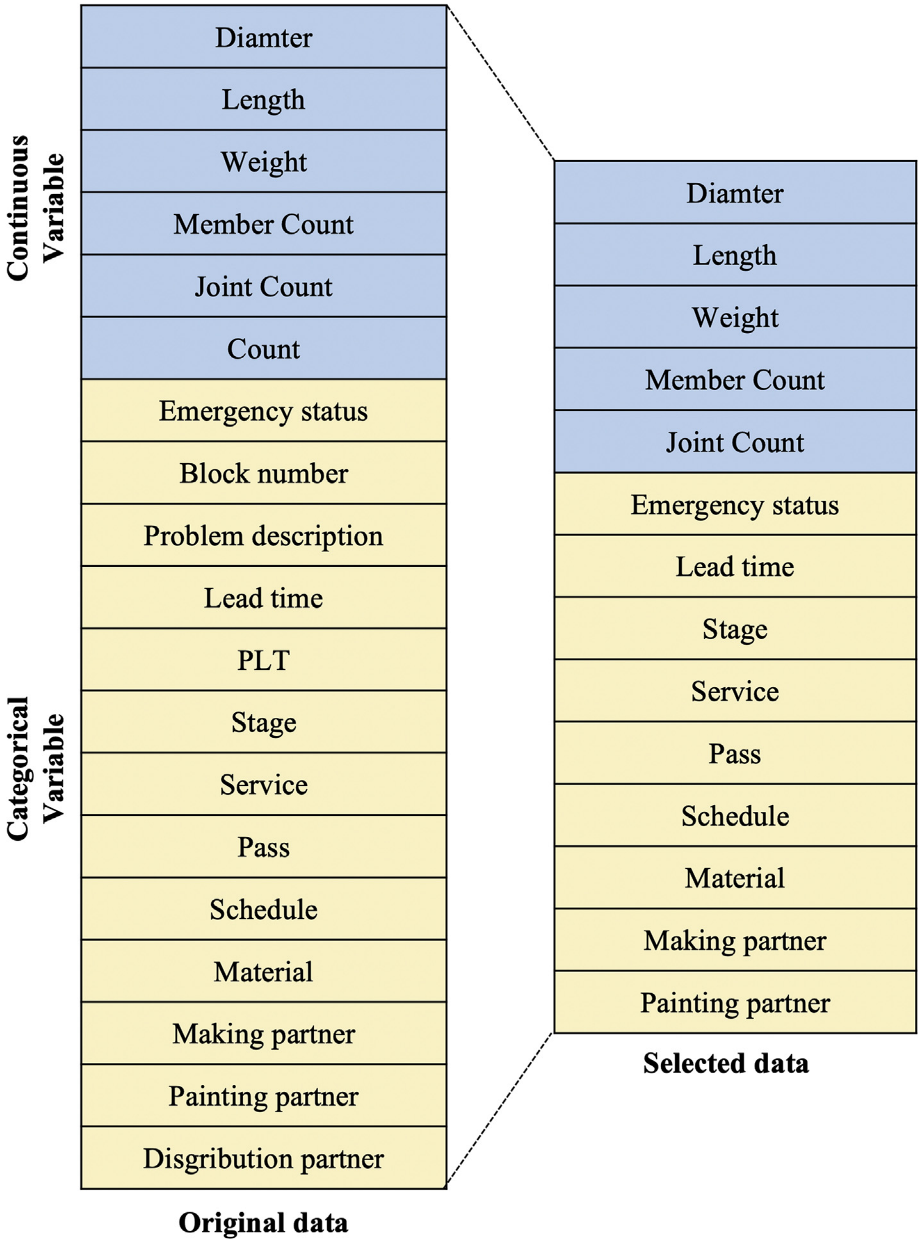


Fig. 8. Independent variables (3).

6.1. Prediction of the lead time for cutting

The first case analyzed in this study is the prediction of the lead time for the cutting of steel using the performance of the cutting process as the training data. Various elements are considered in the planning stage of the cutting process in shipyards. Specifically, seven independent variables were identified and the lead time was defined as a dependent variable by additionally considering the external factors of the process during the analysis of the performance data.

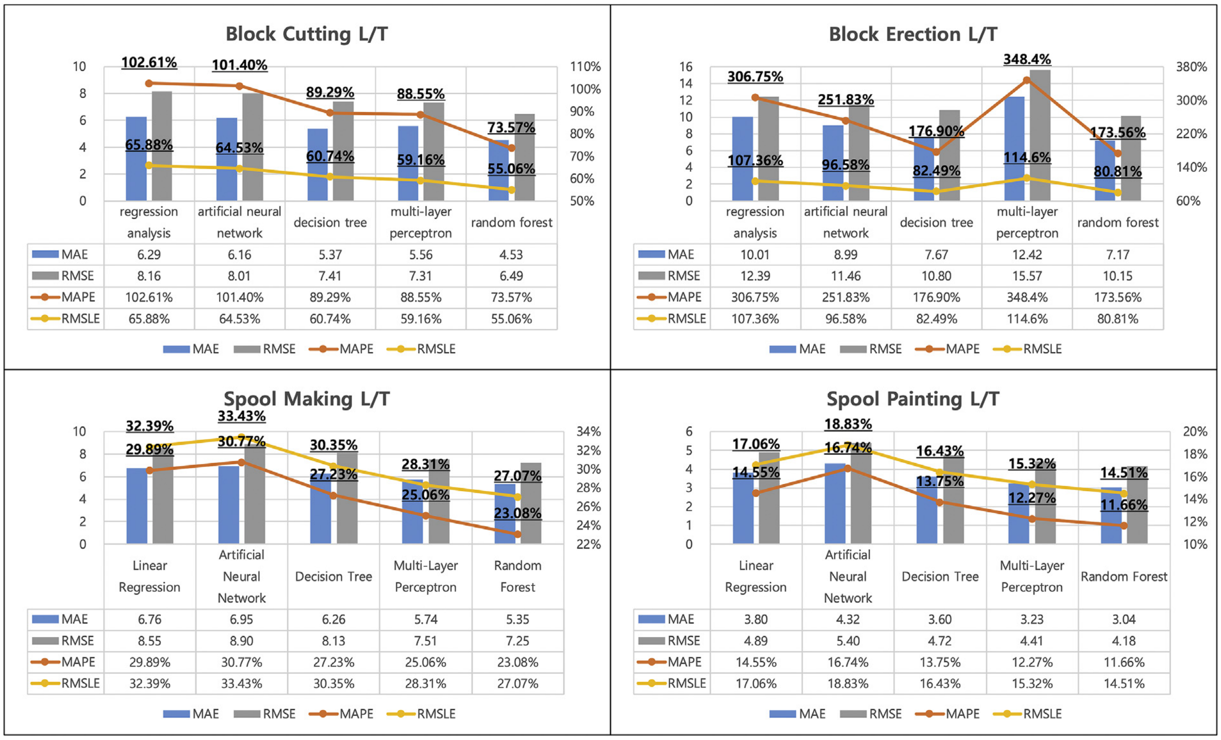
Correlation analysis and the analysis of variables were conducted to analyze the correlations between the defined independent variables and the lead time. Among the first selected variables though interviews with field workers, precipitation was excluded from the independent variables because the correlation coefficient was greater than the reference value of 0.65 as a result of the correlation analysis.

Thus, the analysis was conducted using the six independent variables shown in Fig. 6. Moreover, missing values were simply removed, and outliers were checked and removed by applying the IQR rule and Cook's distance.

6.2. Prediction of the lead time for block erection

The second case is the prediction of the lead time for the erection process of the ship block. The ship block has various data, such as the block code, type, size, and weight. Therefore, it was used as the input data to define 15 independent variables, and the lead time was defined as a dependent variable.

As most of the correlations between the continuous variables were high in the correlation analysis, the independent variables were reduced to decrease the influence of multi-collinearity. In the analysis of variables, it was found that all categorical variables affected the lead time.



Block Cutting L/T

	regression analysis	artificial neural network	decision tree	multi-layer perceptron	random forest
MAE	6.29	6.16	5.37	5.56	4.53
RMSE	8.16	8.01	7.41	7.31	6.49
MAPE	102.61%	101.40%	89.29%	88.55%	73.57%
RMSLE	65.88%	64.53%	60.74%	59.16%	55.06%

Block Erection L/T

	regression analysis	artificial neural network	decision tree	multi-layer perceptron	random forest
MAE	10.01	8.99	7.67	12.42	7.17
RMSE	12.39	11.46	10.80	15.57	10.15
MAPE	306.75%	251.83%	176.90%	348.4%	173.56%
RMSLE	107.36%	96.58%	82.49%	114.6%	80.81%

Spool Making L/T

	Linear Regression	Artificial Neural Network	Decision Tree	Multi-Layer Perceptron	Random Forest
MAE	6.76	6.95	6.26	5.74	5.35
RMSE	8.55	8.90	8.13	7.51	7.25
MAPE	29.89%	30.77%	27.23%	25.06%	23.08%
RMSLE	32.39%	33.43%	30.35%	28.31%	27.07%

Spool Painting L/T

	Linear Regression	Artificial Neural Network	Decision Tree	Multi-Layer Perceptron	Random Forest
MAE	3.80	4.32	3.60	3.23	3.04
RMSE	4.89	5.40	4.72	4.41	4.18
MAPE	14.55%	16.74%	13.75%	12.27%	11.66%
RMSLE	17.06%	18.83%	16.43%	15.32%	14.51%

Fig. 9. Results of learning algorithm.

Thus, the analysis was conducted using the 12 independent variables shown in Fig. 7. Missing values were simply removed and outliers were checked and removed by applying the IQR rule and Cook's distance.

6.3. Prediction of the lead time for the spool procurement process

The third case is the prediction of the lead time for the spool procurement process. In shipyards, the constructions of ships and offshore platforms involve complicated processes from design to production. In the case of offshore platforms, in particular, most of the outfitting processes relate to spools, but problems occur due to delivery delays because proper procurement management for spools is difficult.

The supply chain data of shipyards consist of the time series data of spools by process as well as various data related to spool fabrication and installation. Among them, the lead times of the fabrication and painting processes were targeted in this study. Nineteen independent variables were defined based on the properties of the spools, and the lead times of the fabrication and painting processes were defined as dependent variables.

Correlation analysis and the analysis of variables were conducted to analyze the correlations between the variables. Finally, the analysis was conducted using 14 independent variables, as shown in Fig. 8. Missing values were simply removed and outliers were checked and removed by applying the IQR rule and Cook's distance.

7. Analysis of results of the prediction models

7.1. Machine learning prediction models

Regression analysis, artificial neural network, and decision tree were used as machine learning algorithms, and prediction models were created according to the process data. In this study, prediction was performed by classifying the results of analyzing the raw data into Case 1 and the results of analyzing the preprocessed data into Case 2 to analyze the influence of data preprocessing. Therefore, 24 models were finally created by classifying the data of the four processes, namely, (1) cutting process, (2) erection process, (3) spool fabrication process, and (4) spool painting process, according to the analysis cases. The results of the performance evaluation of the final prediction models are as follows (Tables 3–6)).

Table 6
Results of machine learning (4).

Case	Regression Analysis		Artificial Neural Network		Decision tree	
	1	2	1	2	1	2
MAE	6.01	3.80	6.06	4.32	5.47	3.60
MAPE	35.1%	14.6%	35.5%	16.7%	31.9%	13.8%
RMSE	9.79	4.89	9.68	5.40	9.21	4.72
RMSLE	37.4%	17.1%	37.3%	18.8%	34.8%	16.4%

Table 7
Model cases of MLP.

	Data	Hidden Layer	Batch Size
Case 1	MLP Input Standardization(X)	3	100
Case 2	MLP Input Standardization(X)	3	50
Case 3	MLP Input Standardization(X)	3	30
Case 4	MLP Input Standardization(X)	5	100
Case 5	MLP Input Standardization(X)	10	100
Case 6	MLP Input Standardization(O)	3	100

Table 8
Results of deep learning (1).

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
MAE	5.56	5.59	5.63	5.46	5.64	5.33
MAPE	88.6%	113.0%	100.7%	102.6%	99.6%	96.8%
RMSE	7.31	7.16	7.31	7.01	7.25	6.98
RMSLE	59.2%	66.1%	62.8%	62.8%	62.2%	61.2%

Table 9
Results of deep learning (2).

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
MAE	12.43	12.43	12.43	12.43	12.43	12.43
MAPE	348.7%	348.8%	348.4%	348.9%	348.6%	348.8%
RMSE	15.57	15.57	15.57	15.57	15.57	15.57
RMSLE	114.6%	114.6%	114.6%	114.6%	114.6%	114.6%

Table 10
Results of deep learning (3).

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
MAE	7.71	7.60	7.71	7.62	8.51	5.74
MAPE	33.8%	30.3%	33.6%	34.0%	38.2%	25.1%
RMSE	9.82	10.06	9.90	9.69	10.86	7.51
RMSLE	35.9%	35.2%	36.1%	35.9%	39.6%	28.3%

Table 11
Results of deep learning (4).

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
MAE	4.28	4.33	4.24	4.28	4.29	3.23
MAPE	16.5%	16.8%	16.2%	16.4%	16.6%	12.3%
RMSE	5.38	5.39	5.38	5.38	5.38	4.41
RMSLE	18.7%	18.9%	18.6%	18.7%	18.7%	15.3%

Table 12
Model cases of random forest.

	Ntree	mtry
Case 1	500	5
Case 2	200	5
Case 3	100	5
Case 4	500	3
Case 5	500	10

The analysis using the evaluation criteria confirmed that Case 2 with data preprocessing exhibited higher prediction accuracy for all process data. Among the algorithms, the decision tree model exhibited excellent prediction accuracy. Moreover, the prediction accuracy of the lead time for spool procurement was highest compared to those of the cutting and erection processes, indicating that process variables significantly affected the lead times.

7.2. Deep learning prediction models

The MLP model, which uses the “Keras” library, was applied as the deep learning algorithm. ‘Keras’ provides intuitive APIs for deep learning models, and engines dedicated to deep learning, such as Tensorflow, Theano, and CNTK, are operated internally.

The structure of the MLP model can be defined by setting parameters, such as Epoch (number of learning iterations), Activation (activation function), Batch Size, and Hidden Layer. Model learning, which minimizes the loss function of the deep learning model, can be performed while the weight is updated according to the Epoch and Batch Size.

In this study, Epoch was fixed at 200 times, and the analysis cases were classified according to the number of Hidden Layers and the Batch Size, which is the number of samples used for updating the weight. In addition, the input data were standardized, and the subsequent results of the prediction models were analyzed to examine the influence of the data distribution. Based on the model setting of Case 1, Cases 2 and 3 were defined according to the Batch Size and Cases 4 and 5 were defined according to the number of Hidden Layers. Moreover, Case 6 was defined to examine the influence of the standardization of the input data (Table 7).

The results of the performance evaluation of the final prediction models for the (1) cutting process, (2) erection process, (3) spool fabrication process, and (4) spool painting process are presented in Tables 8–11.

Case 1 with a large Batch Size exhibited a low error rate for the cutting process, and Case 3 with a small Batch Size showed a low error rate for the erection process. In the spool procurement process, Case 6, wherein the input data were standardized, showed the lowest error rate. Because the distribution of the spool fabrication and painting process data exhibited severe skewness, the influence of standardization was considered relatively high. In the above analysis of the cutting and erection processes, the influence of data standardization was considered not significant because relatively even data distribution was observed. Moreover, as the appropriate number of weight updates was different for each data, the Batch Size that ensured a low error rate was different for each data.

Table 13
Results of random forest (1).

	Case 1	Case 2	Case 3	Case 4	Case 5
MAE	4.53	4.53	4.53	4.54	4.54
MAPE	73.6%	73.8%	73.8%	74.3%	73.7%
RMSE	6.49	6.50	6.50	6.43	6.53
RMSLE	55.1%	55.1%	55.1%	54.6%	55.3%

Table 14
Results of random forest (2).

	Case 1	Case 2	Case 3	Case 4	Case 5
MAE	7.26	7.28	7.29	7.47	7.17
MAPE	178.9%	178.2%	179.6%	188.6%	173.6%
RMSE	10.15	10.17	10.20	10.22	10.15
RMSLE	81.6%	81.6%	81.9%	83.6%	80.8%

Table 15
Results of random forest (3).

	Case 1	Case 2	Case 3	Case 4	Case 5
MAE	5.34	5.36	5.37	5.44	5.35
MAPE	23.1%	23.2%	23.2%	23.6%	23.1%
RMSE	7.20	7.22	7.24	7.23	7.25
RMSLE	26.9%	27.0%	27.1%	27.1%	27.1%

Table 16
Results of random forest (4).

	Case 1	Case 2	Case 3	Case 4	Case 5
MAE	3.04	3.06	3.06	3.11	3.04
MAPE	11.7%	11.7%	11.7%	11.9%	11.7%
RMSE	4.18	4.21	4.20	4.21	4.22
RMSLE	14.5%	14.6%	14.6%	14.6%	14.7%

7.3. Ensemble learning prediction models

In this study, a learning model was created using random forest as an ensemble learning algorithm. The structure of the random forest model can be defined by setting parameters such as ntree, which represents the number of decision trees, and mtry, which represents the number of variables to be considered while determining the criterion for dividing the nodes. The default values for ntree and mtry are automatically provided. However, in this study, the values of ntree and mtry were adjusted to further improve the performance of the model, and the analysis cases were classified accordingly. Based on Case 1, Cases 2 and 3 were defined according to the value of ntree and Cases 4 and 5 were defined according to the value of mtry. (see Table 12).

The results of the performance evaluation of the final prediction models for the (1) cutting process, (2) erection process, (3) spool fabrication process, and (4) spool painting process are presented in Tables 13–16.

In all the processes, the error rate was low when the value of ntree was 500. This appears to be because as the value of ntree increased, the generalization error converged to a certain value, which prevented overfitting even though it took longer to construct the model. Moreover, the error rate was low when the value of mtry was 5 in the cutting and spool procurement processes. However, in the case of the erection process, the error rate was low when the value of mtry was 10. Thus, it was confirmed that the appropriate value of mtry was different depending on the data type. When the value of mtry is too small, explanatory variables of small weights are placed in the upper nodes, thereby forming complex trees with increased degree of impurity. When the value of mtry is too large, each tree becomes similar, thereby decreasing the predictive power. Therefore, appropriate tuning is required depending on the data.

7.4. Final analysis results

Various learning algorithms were applied to construct the lead time prediction models for shipbuilding, and the performance of each prediction model was compared. Regression analysis, artificial neural network, and decision tree were used as the machine learning algorithms and MLP was used as the deep learning algorithm. Random forest was used as an ensemble learning algorithm. The results of the performance evaluations of the final prediction models for the cutting process, erection process, spool fabrication process, and spool painting process are given below (Fig. 9).

The analysis of the results according to the process data revealed that the spool fabrication and painting processes showed slightly lower error rates whereas the block erection process exhibited the highest error rate.

Among the machine learning algorithms, the decision tree model exhibited better results than the existing regression analysis and artificial neural network. Moreover, it was confirmed that the models with the deep learning algorithm showed better performance than those with the machine learning algorithms and the models with the ensemble learning algorithm exhibited better performance than those with the deep learning algorithm.

8. Conclusion and future research plan

In this study, the machine learning methodology was applied to systematically establish the master data for production lead times maintained in shipyards. As the lead time master data have high variability in shipbuilding production, the existing engineering methodology has limitations in calculating quantity or time. Therefore, this study attempted to improve the master data, which significantly vary depending on the production environment, by creating production lead time prediction models that consider various product attributes and resources in shipyards.

Open source programming languages, such as R and Python, were used for the data analysis and creation of the prediction models. The prediction models were constructed by applying various learning algorithms available in the development environment. Three types of data were collected from shipyards in this study. The analysis results are as follows.

The first analysis case was the prediction of the lead time for cutting. Performance data for the mid-term schedule of the cutting process were collected and lead time prediction models were created using the algorithms. The average of the Mean Absolute Percentage Error (MAPE) values was 91.1% and that of the Root Mean Squared Logarithmic Error (RMSLE) values was 61.1%. The ensemble learning algorithm exhibited the highest prediction accuracy.

The second analysis case was the prediction of the lead time for block erection. Prediction models were created by identifying the various process variables of the block and using them as independent variables. As a result, the average of the MAPE values was 251.5% and that of the RMSLE values was 96.4%, which were the highest among all the analysis cases. The ensemble learning algorithm exhibited the highest prediction accuracy. In the case of the erection process, it appears that there are limitations to predicting the lead time using the machine learning methodology because the raw data were extremely irregular.

The third analysis case was the prediction of the lead time for spool supply chain. Prediction models were created using analysis algorithms in the same manner as in the above cases. As a result, the lead time for the spool fabrication process had an average MAPE value of 27.2% and the average RMSLE value was 30.3%. For the painting lead time, the average of the MAPE values was 13.8% and that of the RMSLE values was 16.4%. The ensemble learning algorithm exhibited the highest prediction accuracy.

When data preprocessing was intensively performed in the analysis process, the error rates of the prediction models decreased when compared with previous studies. Moreover, it was confirmed that among all lead time prediction models, the prediction models with the ensemble learning algorithm exhibited better performance even though the performance varied depending on the process data. While machine learning and deep learning algorithms exhibit remarkable prediction performance as they are normalized and have large scales, the use of excessively complex models may cause an increase in the generalization error due to overfitting to an insufficient number of training data. Overfitting means that a learning model excessively learns training data. In this case, prediction accuracy above a certain level was observed for the training data, but the model is not accurate for new data. However, in the case of ensemble learning, it appears that better performance was observed because the overfitting problem could be addressed by combining multiple learning results even though learning was performed using a simple algorithm.

These results indicate that the master data can be managed more systematically through the predicted lead times than through the existing standard lead time. Moreover, the predicted lead times can support fast decision-making during work planning and make it possible to gain insight into the analysis technique and variable setting according to the process data.

While machine learning algorithms are commonly used for prediction of values, deep learning algorithms not only predict simple values but also present various methodologies to analyze time series data. Therefore, better decision-making can be achieved during work planning if it is possible to predict the time series for various processes in shipbuilding production in the future.

Acknowledgments

This work was partly supported by National IT Industry Promotion Agency grant funded by Ministry of Science and ICT of Korea (S0604-17-1001, Prediction based smart SCM framework for offshore industry and S1106-16-1020, Production strategy and execution simulation technology development for the optimization of offshore plant production cost). Also, this work was partly supported by Korea Evaluation Institute of Industrial Technology grant funded by the Ministry Of Trade, Industry and Energy of Korea (10050495, Development of the simulation based production management system for the middle-sized shipbuilding companies).

Appendix

To help understand this paper, a description of the variables in the data used in the study was added.

Table 17
Description of the variables in the block cutting process

Column name	Description
Weight (kg)	Weight of the block
Precipitation (mm)	Precipitation in working period
Planning L/T (day)	Lead time planned in production planning step
Ship Type	Type of the ship
Block Group	Group of the block
Block Direction	Position of the block
Planning Cooperation	Block cutting process cooperative company

Table 18
Description of the variables in the block erection process

Column name	Description
Length (m)	Length of the block
Width (m)	Width of the block
Height (m)	Height of the block
Area (m^2)	Area of the block
Sub Weight (ton)	Weight of the block member
Net Weight (ton)	Net weight of the block
Weight (ton)	Net weight of the block plus maximum load weight
Planning L/T (day)	Lead time planned in production planning step
Project No.	Number of the project
Stage	Block erection stage
Division	Block erection section
Construction	Block erection process cooperative company
Block Group	Group of the block
Direction	Position of the block
Block Serial No.	Serial number of the block

Table 19
Description of the variables in the spool procurement process

Column name	Description
DIA	Diameter of the spool
Length	Length of the spool
Weight	Weight of the spool
Member Count	Number of connected members
Joint Count	Number of joints between connected members
Count	Number of the block
Emergency	Priority of spool production
Block	Number of the block
Problem	Encountered problems during the process
Apply Lead Time	Variables related to emergency
PLT	Used Pallet
STG	Spool installation stage
Service	Type of fluid flowing into the spool
Pass	Penetration status
Sch	Thickness of the spool
Material	Material of the spool
Making Co	Spool making process cooperative company
After2 Co	Spool painting process cooperative company
Distribution Co	Spool distribution process cooperative company

References

Ham, D.K., 2016. A Study of Data-Mining Methodology in Offshore Plant's Outfittings Procurement Management. Dissertation, Korea Maritime & Ocean University.

Ham, D.K., Back, M.G., Park, J.G., Woo, J.H., 2016. A study of piping leadtime forecast in offshore plant's outfittings procurement management. J. Soc. Nav. Archit. Korea 53 (1), 29–36.

Hur, M.H., Lee, S.K., Kim, B.S., Cho, S.Z., Lee, D.H., Lee, D.H., 2015. A study on the man-hour prediction system for shipbuilding. J. Intell. Manuf. 26 (6), 1267–1279.

Jo, S.J., Kang, S.H., 2016. Industrial applications of machine learning (artificial intelligence). Ind. Eng. Mag. 23 (2), 34–38.

Jung, S.H., Sim, C.B., 2014. A study on a working pattern analysis prototype using correlation analysis and linear regression analysis in welding BigData environment. J. Korea Inst. Electronic Commun. Sci. 9 (10), 1071–1078.

Kim, S.H., Roh, M.I., Kim, K.S., 2016. A study on big data platform based on Hadoop for the applications in ship and offshore industry. Korean J. Comput. Des. Eng. 21 (3), 334–340.

Lee, Y.H., 2017. A Reference Model for Big Data Analysis in Shipbuilding Industry. Dissertation. Ulsan National Institute of Science and Technology.

Lee, B.W., Yang, J.H., 2008. Ensemble learning of region experts. J. Korea Inf. Sci. Soc. 35 (1A), 120–121.

Lee, J.G., Lee, T.H., Yun, S.R., 2014a. Machine learning for big data analysis. J. Korean Inst. Commun. Sci. 31 (11), 14–26.

Lee, S.K., Kim, B.S., Huh, M.H., Part, J.S., Kang, S.K., Cho, S.Z., Lee, D.G., Lee, D.H., 2014b. Knowledge discovery in inspection reports of marine structures. Expert Syst. Appl. 41 (4), 1153–1167.

Oh, M.J., Roh, M.I., Park, S.W., Kim, S.H., 2018. Estimation of material requirement of piping materials in an offshore structure using big data analysis. J. Soc. Nav. Archit. Korea 55 (3), 243–251.