ICCEPM 2022
The 9th International Conference on Construction Engineering and Project Management
Jun. 20-23, 2022, Las Vegas, NV, USA

Application of Big Data and Machine-learning (ML) Technology to Mitigate Contractor's Design Risks for Engineering, Procurement, and Construction (EPC) Projects

Seong-Jun Choi[1], So-Won Choi[1], Min-Ji Park[1], Eul-Bum Lee[1]

[1]*Graduate Institute of Ferrous and Energy Materials Technology, Pohang University of Science and Technology (POSTECH), Pohang 37673, Korea.* E-mail address: tjdgus947@postech.ac.kr, smilesowon@postech.ac.kr, minjipark@postech.ac.kr, dreblee@postech.ac.kr

Abstract: The risk of project execution increases due to the enlargement and complexity of Engineering, Procurement, and Construction (EPC) plant projects. In the fourth industrial revolution era, there is an increasing need to utilize a large amount of data generated during project execution. The design is a key element for the success of the EPC plant project. Although the design cost is about 5% of the total EPC project cost, it is a critical process that affects the entire subsequent process, such as construction, installation, and operation & maintenance (O&M). This study aims to develop a system using machine-learning (ML) techniques to predict risks and support decision-making based on big data generated in an EPC project's design and construction stages. As a result, three main modules were developed: (M1) the design cost estimation module, (M2) the design error check module, and (M3) the change order forecasting module. M1 estimated design cost based on project data such as contract amount, construction period, total design cost, and man-hour (M/H). M2 and M3 are applications for predicting the severity of schedule delay and cost over-run due to design errors and change orders through unstructured text data extracted from engineering documents. A validation test was performed through a case study to verify the model applied to each module. It is expected to improve the risk response capability of EPC contractors in the design and construction stage through this study.

Key words: Engineering Big Data, Machine-learning (ML), Design Cost Estimation Model, Design Error Check Model, Change Order Forecasting Model

1. INTRODUCTION

In the Engineering, Procurement, and Construction (EPC) contract, the owner wants to complete the construction quickly and take economic profits by operating the facility [1]. As the competition in the global EPC market intensifies, the liability of the EPC contractor increases because the owner intends to pass the project risk to the contractor. The design information provided by the owner at the bidding stage of the EPC plant project is the minimum information for project budget and schedule calculation and consists of schematic drawings and quantity information of the entire facility.

This study aims to support decision-making for risk response when carrying out EPC plant projects. It also seeks to develop a system consisting of a design cost analysis module (M1), design

error analysis module (M2), and design change analysis module (M3) by applying machine learning algorithms based on project data generated during the stages of the EPC project. Each of these three modules is intended to estimate design costs and predict the severity of schedule delay and cost overrun owing to design errors and design changes.

For this study, data from about 43 offshore and 29 onshore EPC projects carried out in the past were collected and analyzed from the perspective of big data to apply machine learning. Pre-processing such as statistical analysis, data refinement, and scaling of collected data was performed to apply the machine learning model suitable for each module. A machine learning model was developed for each module. After that, the performance of each model was verified, and the optimal model was suggested.

2. LITERATURE REVIEW

In the EPC plant project, design is a critical factor for the project's success. Recently, research applying machine learning algorithms, AI, and big data to the EPC industry is increasingly attributed to the extensive and complexation of the EPC industry. Khadtare and Smith [2] advanced a civil construction cost estimation model by using Fractal-COSYSMO system engineering. In addition, Marzouk and Elkadi [3] used an artificial neural network to create a cost estimation model for a water treatment plant, and Pesko and Mucenski [4] applied a machine-learning (ML) algorithm to calculate the civil construction cost estimation model. Elfahham [5] predicted and estimated the construction cost index using neural regression analysis, networks, and time series. However, the plant industry has limitations in that ML, and quantitative statistical analysis is relatively slow compared to other sectors. According to a study by Kaiser [6], design cost accounts for about 5% of the total project cost. Still, the impact on the entire project is very high and affects all processes such as construction and installation, maintenance, and repair. This study presents a model that predicts the accuracy of EPC project design cost using machine learning techniques and also predicts the severity of schedule delays and cost overruns due to design errors or design changes. This study proposes a model to predict the accuracy of design cost by applying ML techniques and to forecast the severity of schedule delay and cost overrun caused by design errors and design changes.

3. RESEARCH PROCESSES

Each of the three modules of this study was conducted as a four-step process. First, data corresponding to each module was collected. Next, we developed a data frame for ML model application through pre-processing. Third, this paper developed a suitable model for each module's objective and recommended an optimal model. Finally, it conducted performance tests for validation. The processes of this study are shown in Figure 1.

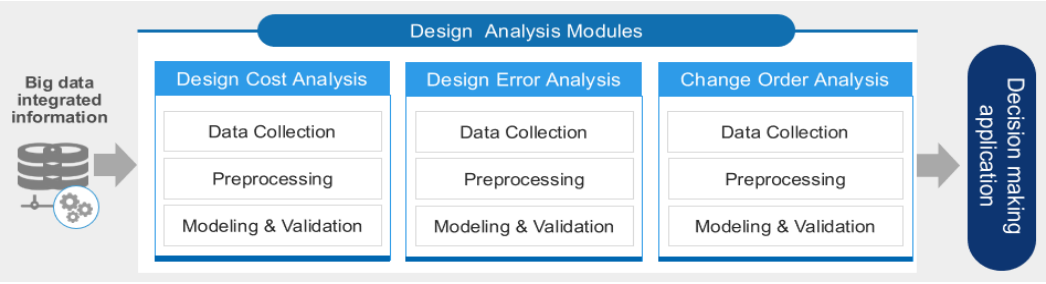

Figure 1. Research Processes

4. MATERIALS & METHODS

4.1. Design Cost Estimation Module

Estimating design cost in an EPC project is one of the important decision-making factors for winning a project as a task performed in the bidding stage, which is the early stage of the project. The design cost analysis module (M1) assists the bidder's decision-making by forecasting design cost through the application of ML techniques. The range of design costs expected in this study is limited to the man-hour required for the design stage.

4.1.1. Data Collection and Pre-processing

For the M1, each construction type has about 1500 project information data such as construction location, contract type, project period, and quantity, including the number of design man-hours and design costs which were collected from 72 EPC projects. These data were organized and converted into a database, as shown in Figure 2 below.

No	Project	Location	Contract	Cost	Duration	Attribute	Type	Structure	Outfiting	Process	Mechanical	Piping	Instrumentation	Design M/H	Design Cost
				1000 USD	M			ton	m2	P&ID	RFQ	Spool	Cable(M)	MH	1000 USD
1	Offshore	Thailand	Lump Sum	1,100,000	27	GAS	Pixed	29544	1410	175	32	17040	488000	295000	24,648
2	Offshore	Malaysia	Lump Sum	820,000	35	GAS	Pixed	24462	2446	87	25	12904	300000	305198	24,718
3	Offshore	Nigeria	Lump Sum	520,000	27	Oil	Pixed	16400	1150	115	29	16850	440670	210500	22,550
4	Offshore	Australia	Lump Sum	800,000	40	Oil	Pixed	29200	6164	168	4	13100	550000	261500	21,570

Figure 2. Composition of Collected Data

Data transformation and feature scaling were performed in the data pre-processing for design cost prediction. Data transformation refers to the process of modifying the structure or format of data to be applied to ML models [7]. First, to use the ML algorithm, string data was converted into integer data, and continuous variables were discretized, normalized, and standardized. Next, feature scaling was performed on numerical variables to improve the ML model's performance. Feature scaling aims to normalize the range of each feature, and it can support the optimal performance of the algorithm [8]. In this module, four scalers were used: Standard Scaler, MinMax Scaler, Robust Scaler, and MaxAbs Scaler. In order to validate the performance of the four scalers, methods such as Scikit Learn and K-fold Cross-Validation were used. The number of Cross-Validation was applied ten times.

As for performance evaluation indicators, $R^2$and Mean Absolute Percentage Error (MAPE) were used. MAPE is an indicator of how much the predicted error rate is used to check the reliability of the regression model [9]. It can be interpreted that the closer the value of R^2 is to 1, the higher the linear correlation is, and the lower the value of MAPE, the smaller the error. As a result of the Validation Result, the Standard Scaler showed the best results. So, a Standard Scaler was used in M1. Table 1 shows the validation result for the four types of scalers.

Table 1. A Validation Result for Four types of Scaling

Scaling Type	Scikit-Learn with Cross-Validation		K-fold Cross-Validation (Fold: 10)	
	[1] R^2	[2] MAPE (%)	R^2	MAPE (%)
Standard Scaler	0.76	21.6	0.7	22.89
MinMax Scaler	0.76	21.6	0.39	35.98
Robust Scaler	0.7	27.42	0.75	21.48
MaxAbs Scaler	0.29	51.25	0.32	38.61

[1] R^2: Coefficient of determination in linear regression analysis [2] MAPE: Mean absolute percentage error

4.1.2. Modeling and Validation

In the M1, since the target variable of design cost data is numerical, a regression analysis algorithm constructed a prediction model. For engineering man-hour prediction, regression analysis algorithms such as Random Forest, Decision Tree, Gradient Boosting, and Elastic Net were applied as prediction models. The evaluation of the prediction model for design man-hour prediction applied two methods (Scikit-Learn and K-fold Cross-Validation) similar to scaling verification.

Since the target variable value of the data to be evaluated is significant, verification was performed by applying the predictive performance evaluation index to MAPE. The MAPE formula for the validation of the Design Cost Prediction Model is shown in Eq.(1).

$$MAPE = \left| \frac{(Predicted\ Cost - Actual\ Cost)}{Actual\ Cost} \right| \times 100\% \quad (1)$$

As a result of Scikit-Learn and K-fold Cross-Validation, Decision Tree showed the lowest prediction error rate. Decision Tree also shows better results in MAPE, although the R^2 value of Random Forest is slightly higher than Decision Tree. Decision Tree with the lowest prediction error rate can be an optimal model with the best performance. Therefore, the decision tree algorithm was recommended for M1. Table 2 shows the validation results of four algorithms for design cost analysis.

Table 2. A Comparison Result of Validation for Design Cost Prediction Algorithm

Applied Algorithm	Scikit-Learn & Cross-Validation with STD Scaler		Cross-Validation (Fold: 10) with STD Scaler	
	[1] R2	[2] MAPE (%)	R^2	MAPE (%)
Random Forest	0.79	17.4	0.85	15.04
Decision Tree	0.77	17.3	0.82	14.39
Gradient Boosting	0.72	20.87	0.84	15.35
Elastic Net	0.76	21.6	0.69	22.17

[1] R^2: Coefficient of determination in linear regression analysis [2] MAPE: Mean absolute percentage error

4.2. Design Error Check Module

The design error analysis module (M2) aims to minimize project risk by predicting the severity of design errors and the severity of schedule delays to determine design errors. Also, we developed an ML-based algorithm that classifies design errors with a high frequency of occurrence, analyzes the severity of design errors according to their causes, and predicts the severity of schedule delays.

4.2.1. Data Collection and Pre-processing

For the M2, approximately 9000 design error data such as design error type, crash report, cause of design error, and project information were collected from 72 EPC projects over the past 20 years. In particular, the design error reference data is a design drawing error of the information providing EPC company or data standardized based on specifics corresponding to the entire design cycle, including details pointed out by the owner during the 3D Modeling Review (30%, 60%,

90%). In addition, to classify the severity of design errors, standardized data were referred to and classified into error causes and error types, design delay severity, and design error severity, as shown in Table 3.

Table 3. Classification of Design Error Type and Severity

Class	Sub-Class		Description
Design Delay Impact	shorter than 26 months	Safe	$0 \leq$ Delay Days < 2
		Marginal	$2 \leq$ Delay Days ≤ 15
		Serious	$15 <$ Delay Days
	longer than 26 months	Safe	$0 \leq$ Delay Days < 3
		Marginal	$3 \leq$ Delay Days ≤ 25
		Serious	$25 <$ Delay Days
Design Error Impact	A	Safe	Simple modification within Scope
	B	Marginal	Cost Impact for 2 Disciplines
	C	Serious	Cost Impact for more than 2 Disciplines

In the M2, the data pre-processing proceeded with feature scaling and feature selection. First, feature scaling was performed to adjust the distribution of data values to improve the ML model's performance for design error analysis. In this module, data cleansing was carried out, by changing data structures, correcting missing values, removing duplicate values, removing outliers, and linking data. As in the design cost analysis module above, the data was scaled through four scalers. The scaler showing optimal performance was selected.

Because the reason for design errors is unstructured data, it requires feature selection for analysis. In order to select features from text variables, text data formalization and integer vectorization must be preceded in advance [10]. Formalization was performed by extracting headwords, tokenization, and stopword processing, and as a result, text data of design errors were converted into data usable for analysis. In addition, unbalanced data were rearranged through integer vectorization. In this study, integer vectorization means that the pre-processed English and Korean text data are arranged in high frequency to low frequency, and then integers are assigned first from low-frequency words to generate vectors. In this way, principal component analysis was applied to many integer vectors to select features from text variables, such as the reason for design errors.

4.2.2. Modeling and Validation

In the M2, design error data is a categorical variable. So, the model was constructed using classification analysis algorithms such as Random Forest, Decision Tree, XGBoost, and Gradient Boosting. Also, prediction and classification algorithms were used as reference models. Because a predictive model is a supervised learning model like a classification model, it learns from labeled training data. Still, unlike a classification model, it uses labeled training data to express the correlation between features and labels as a function [11].

The M2 was tested by applying four classification algorithms to predict delay severity and design error severity due to design errors. ML models were tested using a standard scaler. Of the total data, 80% was used as training data, 20% of data was used as evaluation data, and 10-fold Cross-Validation was used. In addition, to evaluate the performance of the design error analysis model, the F-measure measurement method using the harmonized average of precision and recall

was used. The F-measure evaluation method is mainly used for performance evaluation of ML using classification algorithms such as design error and design change models [12].

Table 4 below shows the results of comparing the performance of classification models for design error analysis. As a result of 10-fold Cross-Validation, Random Forest recorded the highest detection accuracy with an F-measure value of 53%, while all four models did not show a significant difference in F-measurement values. Although the accuracy prediction rate of the four models is higher than 50%, it is necessary to improve the performance through further studies.

Table 4. Validation Results for 4 Types of Design Error Analysis Model

Testing Models	Applied Algorithm	Cross-Validation with STD Scaler (Fold: 10)		
		Precision (%)	Recall (%)	F-measure (%)
Prediction Model #1	Random Forest	53%	53%	53.0%
Prediction Model #2	Decision Tree	52%	50%	51.0%
Prediction Model #3	Gradient Boosting	51%	50%	50.5%
Prediction Model #4	XGBoost	51%	53%	52.0%

4.3. Change Order Forecasting Module

The change order analysis module (M3) is a module for predicting schedule severity and cost severity using change order information. Design error analysis and change order analysis are fundamentally peer analysis methods in that both design errors and design changes predict the impact on schedule and cost. As a result, semi-structured data, application of the same data pre-processing technique, and algorithms for analysis also applied the same series. However, the data sets are slightly different due to the details of the collected data.

4.3.1. Data Collection and Pre-processing

For the M3, about 3000 change order data such as change order report, the reason for design change, design change type, revision history of P&ID and plot plan were collected from 72 EPC projects in the past. The most important among various data items is the reason for design change, and it is classified into change type, schedule severity, and cost severity, as shown in Table 5. The severity of the schedule delay was classified as the ratio of the design period to the total project cost. The severity of design change cost overrun was classified according to the ratio of the total design change amount to the total project cost.

Table 5. Classification of Design Change type, Schedule Delay Severity and Cost Overrun.

Class		Sub-Class
Schedule Delay Impact due to Design Change	Safe	$0 \leq$ Delay of Total Schedule $< 1\%$
	Marginal	$1 \leq$ Delay of Total Schedule $< 2\%$
	Serious	$2\% \leq$ Delay of Total Schedule
Cost Overrun Impact due to Design Change	Safe	$0 \leq$ Cost overrun of Total Cost $< 5\%$
	Marginal	$5\% \leq$ Cost overrun of Total Cost $< 10\%$
	Serious	$10\% \leq$ Cost overrun of Total Cost

In the M3, pre-processing was carried out similarly to the M2. The data was done through the four scalers mentioned above. The scaler showing optimal performance was selected, and data that did not affect the analysis such as design ID were deleted using regular expressions. All needless

variables were removed and text data such as design change reasons was pre-processed. After all, feature selection was performed through integer vectorization of text data and formalization.

4.3.2. Modeling and Validation

The M3 constructed an ML model using a classification algorithm similar to design errors analysis because a target variable for design change data is a categorical variable. The causes of design change were classified, and the severity of schedule delay and cost overrun was analyzed through a predictive model based on design change history data.

The M3 used four classification algorithms like design error analysis, and 10-fold Cross-Validation was used for the applied model. In addition, the test was conducted using data scaled with a standard scaler, and the training data and verification data were divided and applied at a ratio of 80:20. The evaluation was made using the F-measure.

As a result of Cross-Validation, Random Forest recorded the highest design change detection accuracy with an F-measure value of 66.5%, and Decision Tree showed the lowest prediction rate with an F-measure of 57.4%. Although all four models provide a prediction rate of more than 50%, further improvement is needed to advance performance. Table 6 shows the validation results of four predictive models.

Table 6. Validation Results for 4 Types of Change Order Analysis Model

Testing Models	Applied Algorithm	Cross-Validation with STD Scaler (Fold : 10)		
		Precision (%)	Recall (%)	F-measure (%)
Prediction Model #1	Random Forest	66%	67%	66.5%
Prediction Model #2	Decision Tree	55%	60%	57.4%
Prediction Model #3	Gradient Boosting	60%	63%	61.5%
Prediction Model #4	XGBoost	64%	60%	61.9%

5. CONCLUSIONS

The purpose of this study is to respond to project risks based on data generated during the design and construction phase of the EPC project and to help engineers make decisions. First, for the analysis of the EPC project, design information such as project data of about 72 EPC plant projects, including the contract price and design cost, and the owner's comments on 2D Modeling and design error reports were collected. The collected data was databased, and data pre-processing was performed to apply it to the ML model. Through this study, the design cost analysis module, design error analysis module, and change order analysis module were developed, and an ML model was developed for each module.

The M1 aims to predict design estimates by analyzing man-hour input costs to engineers, and the M2 aims to predict the severity of design errors and the resulting schedule delay severity. Lastly, the purpose of the M3 is to predict the severity of schedule delays and cost overruns due to design change by analyzing the cause of design change.

The M1 is 14.39% MAPE of the Decision Tree model, and the design cost prediction accuracy of 85% was confirmed. In the M2 and M3, Random Forest showed the highest prediction rates with F-measure values of 53% and 66.5%, respectively. Currently, the accuracy of the prediction rate of the M2 and M3 is low, but performance improvement can be expected through the accumulation of more data in the future. It is hoped that the use of three design modules based on objective and quantitative evidence reflecting the characteristics of the project will be increased for engineering practitioners.

REFERENCES

[1] D. McNair, “EPC Contracts in the Power Sector, Asia Pacific Projects Update”, DLA, 2011.

[2] M. Khadtare, E. Smith, “Fractal-COSYSMO Systems Engineering Cost Estimation for Complex Projects”, Procedia Computer Science, 2011.

[3] M. Marzouk, M. Elkadi, “Estimating water treatment plants cost using factor analysis and artificial neural networks”, Journal of Cleaner Production, vol. 112, pp. 4540-4549, 2016

[4] I. Pesko, V. Mucenski, “Estimation of Costs and Durations of Construction of Urban Roads Using ANN and SMV”, Complexity, vol. 2017, pp. 1-13, 2017.

[5] Y. Elfahham, “Estimation and prediction of construction cost index using neural networks, time series, and regression”, Alexandria Engineering Journal, vol. 58, no. 2, pp. 499-506, 2019.

[6] M. J. Kaiser, “The Offshore Pipeline Construction Industry”, Elsevier Science & Technology, pp. 229-253, 2020.

[7] S. Sajid, V. Marius, A. Soylu, D. Roman, “Predictive Data Transformation Suggestions in Grafterizer Using Machine Learning”, Communications in Computer and Information Science, pp. 137-149, 2019.

[8] T. Li, B. Jing, N. Ying, “Adaptive Scaling”, Cornell University, 2017. https://arxiv.org/abs/1709.00566

[9] A. De Myttenaere, B. Golden, B. Le Grand, F. Rossi, “Mean Absolute Percentage Error for regression models”, Neurocomputing (Amsterdam), vol. 192, pp. 38-48, 2016.

[10] T. Mikolov, W. Yih, G. Zweig, “Linguistic regularities in continuous space word representations”, In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013), Atlanta, USA, pp. 746-751, 2013.

[11] Z. Yu, M. Zhang, “Multi-Label Classification with Label-Specific Feature Generation: A Wrapped Approach”, IEEE Electronic Library (IEL) Journals, 2021.

[12] J-M. Lee, “Developing Cyber Risk Assessment Framework for Cyber Insurance: A Big Data Approach”, Insurance Research Institute, Vol. 2018, no. 15, pp. 1-80, 2018.