CCC 2018

Proceedings of the Creative Construction Conference (2018)
Edited by: Miroslaw J. Skibniewski & Miklos Hajdu
DOI 10.3311/CCC2018-113

Creative Construction Conference 2018, CCC 2018, 30 June - 3 July 2018, Ljubljana, Slovenia

Development of the prediction model of workers with fatal accident at construction site using machine learning

Jongko Choi[a], Bonsung Gu[b], Sangyoon Chin, Ph.D[c,*]

[a] *Department of Convergence Engineering for Future City, Sungkyunkwan University, Suwon, 16419, S. Korea*
[b] *Department of Industrial Engineering, Sungkyunkwan University, Suwon, 16419, S. Korea*
[c] *Professor, School of Civil, Architectural Engineering & Landscape Architecture, Sungkyunkwan University, Suwon, 16419, S. Korea*

Abstract

In the Republic of Korea, the industrial accident rate and the number of casualties in the construction industry have continued to increase for six years from 2011. The average industrial accident rate is 26.44% and the average number of casualties is 24,183. To prevent accidents, the Ministry of Employment and Labor (MOEL) presents various analysis data through the annual industrial accident report, but this has not been effective in reducing accidents as a result. In this paper, using the Logistic regression model that is one of the Machine learning method, this study develops a construction accident prediction model by training 80% of the data of accident casualties (25,114 persons) and accident deaths (499 persons) by 2016 and tests the predicted model with 20% unused data. And then, this study presents a construction site safety management process using the predicted model. The model and process developed in this paper are expected to contribute to the safety management of the construction site as a tool to prevent fatal accidents of construction workers.

Peer-review under responsibility of the scientific committee of the Creative Construction Conference 2018.

Keywords: Demographic Characteristics, Machine Learning, Public Data, Safety Management, Prediction of fatal Accident

1. Introduction

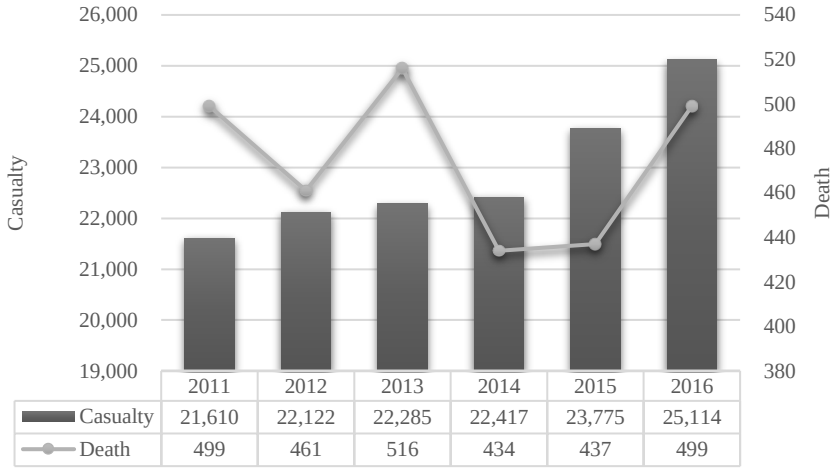
1.1. Research background and purpose

When an industrial accident happens in Korea, accidents are reported to the Korea Workers' Compensation and Welfare Service (COMWEL). COMWEL determines whether an industrial accident is registered and collects information about the accident victims. This data is shared with MOEL, which makes up the industrial accident report. In this report, the cumulative value of survey items is presented, and it is reflected in the safety management with reference to the construction site. According to the report of MOEL, the number of industrial accidents had increased every year for six years from 2011 to 2016 (Figure 1), an annual average of 23,362 injured people, and 2,846 people died. This represents an annual average of 26.44 % of the total number of accident victims in the entire industry (MOEL, 2012; 2013; 2014; 2015; 2016; 2017; Choi et al. 2017). The result means that the statistical data did not influence the accident prevention of construction workers. It needs an easy way to use statistical data in the field. Therefore, the purpose of this paper is to propose a useful model to prevent fatal accidents in the field by utilizing industrial accident data collected from Korean public institutions.

Construction accidents are caused by physical causes (unsafe conditions, 10%), human causes (unsafe behavior, 88%), and indirect causes (environmental causes, 2%) (Heinrich et al. 1980). However, on-site training is conducted

Corresponding author: Author email: schin@skku.edu

to eliminate unsafe behaviors that directly cause disasters, and there is no consideration of personal defects or personal characteristics, which are potential indirect causes of direct causes (Kim, 2008). This study classifies the workers who have a high probability of fatal accidents (death) by Machine-learning (Logistic regression model).



	2011	2012	2013	2014	2015	2016
Casualty	21,610	22,122	22,285	22,417	23,775	25,114
Death	499	461	516	434	437	499

Figure 1. Number of casualties and deaths by year (2011-2016)
(MOEL, 2012; 2013; 2014; 2015; 2016; 2017)

1.2. Research scope

The scope of the analysis was limited based on 2016 industrial accident data collected by MOEL based on age, sex, duration of career, accident date, project size, classification of casualty and death, and construction type. In 2016, industrial accidents consist of 25,114 casualties and 499 deaths.

1.3. Method

Machine learning about industrial accident information possessed by a public organization creates a predictive model that evaluates the risk of fatal accident when construction workers get to work. Depending on the degree of risk assessed, the safety manager can efficiently manage the workers based on the information of the day. First, collect construction workers' data from MOEL. Second, the collected data (age, sex, duration of career, project size, construction type, accident date) is pre-processed. Third, derive a regression equation that predicts fatal accidents (death) by machine-learning industrial accident data from a Logistic regression model.

2. Development of the prediction model for fatal accident

2.1. Data preprocessing

Data was received from MOEL through a public information portal (www.open.go.kr). It is personal information about the casualties and deaths of 2016. Values categorized in the supplied data require conversion for data analysis. For categorical data, the median was taken as the reference value. The unit of duration of career was converted into months.

2.2. Logistic regression model

D. R. Cox (1958) proposed a Logistic regression model. It is a statistical technique used to predict the likelihood of an event using a linear combination of independent variables as a probability model (Hosmer, D.W. et al. 2013). The

Logistic regression is widely used in a variety of fields, including construction or social analysis. Wong (2004) used the Logistic regression model to formulate the Contract Execution Index for UK contractors. The goal of Logistic regression is the same as the goal of the general regression analysis. The relationship between the dependent variable and the independent variable is expressed as a linear combination function and used in future prediction models (Aram So et al. 2017). However, unlike linear regression analysis, Logistic regression can be seen as a kind of classification technique because the dependent variable includes categorical data and when the input data is given, the result of the data is divided into specific categories.

This study modeled it on the basis of a Logistic model because the data used in this paper included categorical data. Also, since it can be expressed in a simple line form, field managers can easily use the equation obtained based on this Machine learning method. The basic approach of Logistic regression is to use linear regression. The linear prediction function can be expressed as follows for a particular data term (Equation 1).

$$f(i) = \beta_0 + \beta_{1}x_{1i} + \cdots + \beta_{m}x_{mi} \quad (1)$$

The Logistic model expression ensures that the dependent variable or result value is between the probability value [0, 1], as shown in Figure 2, regardless of the number of independent variables [-∞, ∞]. The result is classified as 0 or 1 according to the value 0.5.

In this paper, 80% of the total data was sampled during Machine learning, and it was verified as 20%. Since the data of the accident is 50 times more than the data of death, the data of death to predict the occurrence of a fatal accident is weight-learned.

2.2.1. Application of logistic regression model

First, the entire data was applied to the Logistic regression model (Table 1). As a result of the analysis, the age, duration of career, project size, whether it was Sunday or Thursday, and Architecture or Railway or track construction type or not influence on fatal accident. Second, the Logistic regression model was reapplied except for the data with insignificant impact. As a result, Table 2 was derived, and thus Equation 2 was determined. According to Table 2, the probability of occurrence of fatal accidents increases with old age, large project size, long duration of career, on Sunday, and in April. Also, the probability of occurrence in case of architecture project is lowered. A system using Equation 2 can easily classify workers who are likely to have a fatal accident.

Table 1. Apply the whole data to the Logistic regression model

		Estimate Std.	Error	z value	Pr(>|z|)	
(Intercept)		-4.406.E+00	4.329.E-01	-1.018.E+01	< 2e-16	***
Age		1.583.E-02	4.777.E-03	3.313.E+00	9.240.E-04	***
Woman		-3.528.E-01	3.410.E-01	-1.035.E+00	3.008.E-01	
Project size		1.159.E-03	1.639.E-04	7.072.E+00	1.530.E-12	***
Duration of career		7.867.E-03	1.805.E-03	4.357.E+00	1.320.E-05	***
Day of the week	Sunday	4.324.E-01	1.803.E-01	2.399.E+00	1.646.E-02	*
	Monday	-2.309.E-01	1.768.E-01	-1.306.E+00	1.917.E-01	
	Tuesday	-1.789.E-02	1.680.E-01	-1.060.E-01	9.152.E-01	
	Wednesday	9.085.E-03	1.680.E-01	5.400.E-02	9.569.E-01	
	Thursday	2.596.E-01	1.574.E-01	1.650.E+00	9.902.E-02	.
	Saturday	6.414.E-02	1.691.E-01	3.790.E-01	7.045.E-01	
Month	February	1.782.E-01	2.647.E-01	6.730.E-01	5.008.E-01	
	March	7.928.E-03	2.485.E-01	3.200.E-02	9.746.E-01	
	April	4.548.E-01	2.327.E-01	1.955.E+00	5.060.E-02	.
	May	-6.366.E-02	2.522.E-01	-2.520.E-01	8.008.E-01	
	June	4.328.E-02	2.420.E-01	1.790.E-01	8.581.E-01	
	July	2.004.E-01	2.385.E-01	8.400.E-01	4.008.E-01	
	August	-9.342.E-02	2.483.E-01	-3.760.E-01	7.068.E-01	
	September	8.261.E-02	2.500.E-01	3.300.E-01	7.411.E-01	
	October	2.036.E-02	2.430.E-01	8.400.E-02	9.332.E-01	
	November	-9.222.E-02	2.508.E-01	-3.680.E-01	7.131.E-01	
	December	-5.375.E-02	2.537.E-01	-2.120.E-01	8.322.E-01	
Const -ruction Type	Architecture	-7.539.E-01	2.939.E-01	-2.565.E+00	1.031.E-02	*
	Highway and subway	-1.043.E+01	5.354.E+02	-1.900.E-02	9.845.E-01	
	(High) dam	-1.015.E+01	3.086.E+02	-3.300.E-02	9.738.E-01	
	Machinery	1.717.E-01	3.846.E-01	4.460.E-01	6.553.E-01	
	Road	-1.426.E-01	1.069.E+00	-1.330.E-01	8.939.E-01	
	Hydropower Facility	-1.059.E+01	5.354.E+02	-2.000.E-02	9.842.E-01	
	Railway or track	1.557.E+00	8.172.E-01	1.906.E+00	5.669.E-02	.
	Others	-4.062.E-01	2.964.E-01	-1.371.E+00	1.705.E-01	

Table 2. Apply the data determined to be correlated to the Logistic regression model

	Estimate Std.	Error	z value	Pr(>|z|)	
(Intercept)	-4.644.E+00	3.013.E-01	-1.541.E+01	< 2e-16	***
Age	1.483.E-02	4.741.E-03	3.128.E+00	1.760.E-03	**
Project size	1.164.E-03	1.634.E-04	7.127.E+00	1.030.E-12	***
Duration of career	8.262.E-03	1.762.E-03	4.688.E+00	2.760.E-06	***
Sunday	4.071.E-01	1.443.E-01	2.821.E+00	4.790.E-03	**
April	4.256.E-01	1.402.E-01	-3.037.E+00	2.390.E-03	**
Architecture	-4.097.E-01	9.478.E-02	4.323.E+00	1.540.E-05	***

$$f(risk) = 1.483e^{-2}x_1 + 1.164e^{-3}x_2 + 8.262e^{-3}x_3 + 4.071e^{-1}x_4 + 4.256e^{-1}x_5 - 4.097e^{-1}x_6 - 4.644 \quad (2)$$

Note: χ_1 = Age, χ_2 = Project size, χ_3 = Duration of career, χ_4 = Day of the week (Sunday = 1, except for that = 0), χ_5 = Month (April = 1, except for that = 0), χ_6 = Construction type (Architecture = 1, except for that = 0) ※ Only one construction type must be selected.

2.2.2. Logistic regression model performance

In this paper, the performance is verified using k-fold cross validation (Kohavi, R., 1995). This is a method for increasing the statistical reliability of the classifier performance measurement in the machine learning field. The method is as follows. Divide the sample into groups of k. k-1 sets train the classifier and the other set to measure the performance of the classifier apply to the model. This process can be performed k different times, and the accuracy of the acquired k times can be averaged and defined as the performance of the classifier. The advantage of cross validation is that you can use most of the samples you have in Training. In the experiment, 5-fold crossover verification was conducted with 20 % of the data after learning from 80 % of the data.

The data in this paper are severely imbalanced, with the accident victim being 50 times more likely than the death victim. If the study calculates a typical error by creating a model that predicts all the predictors and all of the deaths fail, the study will have about 98 percent of performance. AUROC (Davis, J. et al. 2006) is a common means of measuring imbalance data performance in the field of machine learning. The closer to 1 is better the performance than the closer to 0.5. The result of AUROC measurements are shown in Figure 3. AUROC means the lower part of the graph line, and 0.6636 is simply a numerical value.

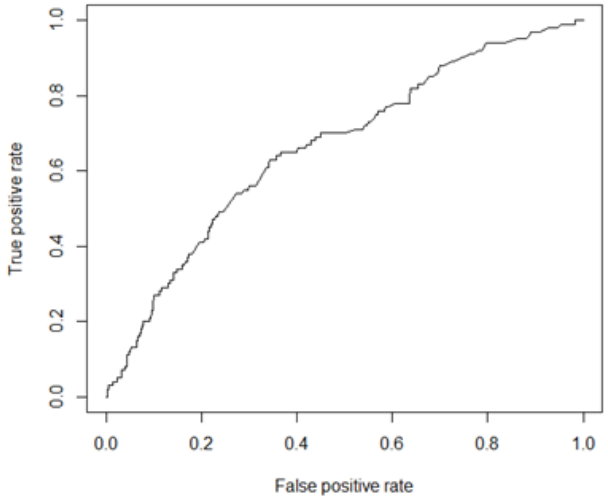


Figure 2. Result of AUROC measurement

3. Conclusion

This paper analyzed the data of 25,613 industrial accident victims of construction industry in Korea (25,114 accident casualties, 499 accident deaths). This study modeled the data based on a Logistic regression model because the data used in this paper included categorical data. For categorical data, the median was taken as the reference value. The unit of duration of career was converted into months. 80% of the total data were sampled and learned, and then 20% were verified. Since the data of casualties are 50 times more than the data of deaths, the weight of deaths is weight-learned in order to predict the occurrence of fatal accident in case of industrial accidents. All elements were analyzed and the Logistic regression model was applied again as a factor affecting the prediction. In this process, First, the sex was excluded. Second, the remaining days except Sunday and Thursday were excluded. Third, the month except April was excluded. Fourth, the construction types except for architecture and Railway or track were excluded. As a result, Equation 2 was derived. In the case of fatal accident, the probability of occurrence is higher when the age is higher, the duration of career is longer, the project size is larger, on Sunday, in April, and the probability of occurrence in case of architecture project is lowered. Finally, the 5-fold cross-validation of the AUROC model was 0.6636. Therefore, when the RFID, QR code, and biometric information are input to the device confirming the worker's work, the risk index is calculated and reported. Based on this index, the safety manager can classify and manage the daily risk groups efficiently, which is expected to prevent fatal accidents (Figure 4).

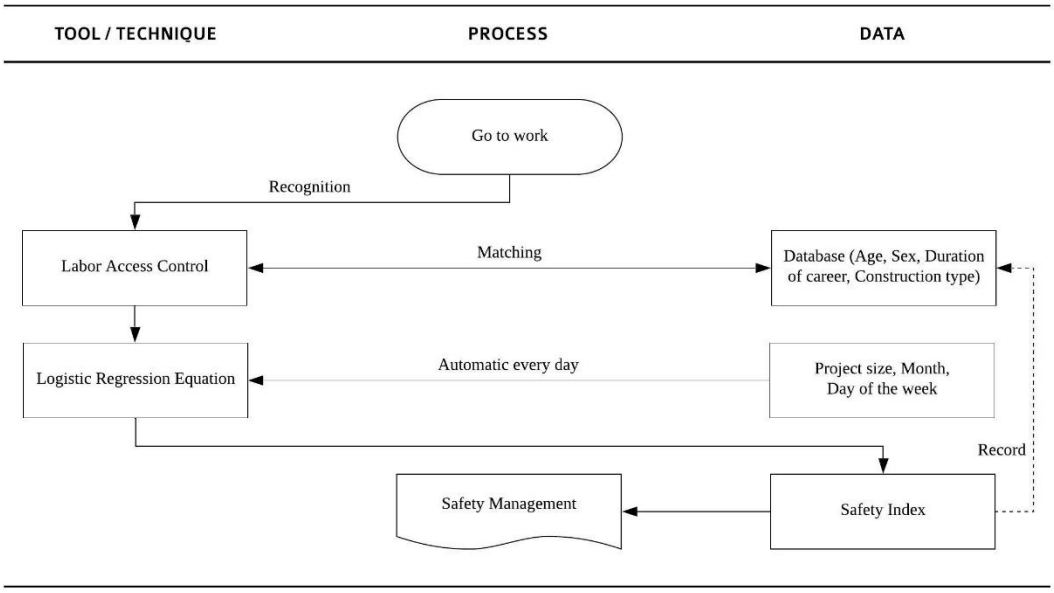


Figure 3. Safety management process using safety index

4. Limitation and Future Research

Currently, the data used in this paper is limited to 2016. By collecting yearly accident data and performing Logistic regression analysis, a result can obtain a higher performance than the Logistic regression equation derived from this paper. If data on general workers who are not hurt are collected in addition to those of the victims, it will be possible to carry out studies capable of predicting accidents as well as fatal accidents.

Acknowledgements

This work is financially supported by Korea Ministry of Land, Infrastructure and Transport (MOLIT) as 「Smart City Master and Doctor Course Grant Program」

References

[1] Aram So, Danial Hooshyar, Kun Woo Park, Heui Seok Lim (2017). Early Diagnosis of Dementia from Clinical Data by Machine Learning Techniques, Applied Sciences, Vol. 7, 651.
[2] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological), 215-242.
[3] Choi, J., Lee, M., Chin, S. (2017). Basic research for analyzing demographic characteristics and industrial accident relation of construction workers and development of on-site safety management process, Proceedings of KICEM Annual Conference, 2017, Vol. 16, 91-92.
[4] Davis, J., Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning, 233-240.
[5] HEINRICH, H. W., PETERSEN, D. C., ROOS, N. R., HAZLETT, S. (1980). Industrial accident prevention: A safety management approach, McGraw-Hill Companies.
[6] Hosmer, D.W., Jr, Lemeshow, S., Sturdivant, R.X. (2013). Applied Logistic Regression, John Wiley & Sons: Hoboken, NJ, USA, Volume 398.
[7] Kim, E. (2008). A Model for Applying Methods of Safety Education Reflecting Individual Properties of Construction Workers, Doctoral thesis, Ajou University, 3
[8] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai, Vol. 14, No. 2, 1137-1145.
[9] Ministry of Employment and Labor. (2012). Analysis of industrial accidents in 2011, 29
[10] Ministry of Employment and Labor. (2013). Analysis of industrial accidents in 2012, 29
[11] Ministry of Employment and Labor. (2014). Analysis of industrial accidents in 2013, 29
[12] Ministry of Employment and Labor. (2015). Analysis of industrial accidents in 2014, 41, No. 11-1490000-000022-10.
[13] Ministry of Employment and Labor. (2016). Analysis of industrial accidents in 2015, 41
[14] Ministry of Employment and Labor. (2017). Analysis of industrial accidents in 2016, 41.
[15] Wong, C. H. (2004). Contractor performance prediction model for the United Kingdom construction contractor: Study of logistic regression approach. Journal of construction engineering and management, 130(5), 691-698.