Developing predictive models of construction fatality characteristics using machine learning

Jianbo Zhu[1], Qianqian Shi[2*], Qiming Li[1], Wenchi Shou[3], Haijiang Li[4], Peng Wu[5]

1. School of Civil Engineering, Southeast University, Nanjing 211189, China
2. College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
3. School of Engineering, Design and Built Environment, Western Sydney University, Sidney, NSW 2751, Australia
4. Cardiff School of Engineering, Cardiff University, Cardiff CF24 3AA, U.K
5. School of Design and the Built Environment, Curtin University, Bentley 6102, Western Australia, Australia

*Corresponding author

Abstract

Construction fatalities have significant economic and emotional burdens to construction employees, families, and organizations. Understanding critical factors influencing construction fatalities and eventually developing predictive models to predict construction fatality characteristics are therefore important. Such activities, which are traditionally based on questionnaire and simple statistical analysis, can now be conducted using comprehensive datasets on construction fatality and advanced machine learning approaches. This study aims to develop predictive models of construction fatality characteristics, including nature of injury (NOI), part of body (POB), source of injury (SOI), and event or exposure (EOE) using machine learning approaches. 30 explanatory variables from 694 fatalities reported by the National Institute for Occupational Safety and Health are used to build the predictive models, with prediction accuracy of 56.6%, 54.0%, 76.5% and 84.9% for NOI, POB, SOI, EOE respectively. Specifically, the model has a prediction accuracy of 84.7% for construction fall fatalities. Important indicators for predicting SOI and EOE are largely the same, with the most important ones being the likelihood of fall, PFAS (functionality and relevant training), workers' activity, onsite safety equipment and install safety protection. Similarly, important indicators for predicting NOI and POB include fall, PFAS, injury year, workers' activity, location and safety equipment. The results will offer useful guidance for construction organizations to establish relevant emergency response plans and first aid facilities and services that correspond to the most likely NOI, POB, SOI and EOE on construction sites.

Keywords: Machine learning; Construction Safety; Fatality; Safety management

1. Introduction

Health and safety issues are ongoing concerns for the construction industry. The construction industry employs a significant number of workforce and fatal injuries are very common. In the United States, 1,061 worker fatalities happened in construction in 2019, which is about 20% of worker fatalities in the private industry (U.S. Department of Labor, 2022). Similarly, in the UK construction industry, fatal injuries happen at a statistically higher rate in construction (Health and Safety Executive, 2022). It is found that the fatality rate for workers in construction is 4 times of all industry average. Given the importance of health and safety issues for construction workers, it is critical to identify, analyze and understand contributing factors to establish relevant injury prevention strategies (Dong et al., 2017). Many studies have therefore been initiated to achieve this objective.

A large collection of studies focuses on investigating factors influencing unsafe behaviors using questionnaires and factor analysis. For example, Yu et al. (2014) identified five categories of key factors, including safety attitude, site safety, government supervision, market restriction and task unpredictability through a questionnaire survey of 104 participants. In recent years, researchers have started to use more advanced data analysis methods to evaluate safety indicators and predict safety performance in construction projects. For example, Xia et al. (2018) used the Bayesian network approach to identify and analyze influencing factors of safety performance through a survey of 142 participants in China. Guo et al. (2016) used structural equation modelling to predict safety behavior of construction workers through a survey of 213 participants in New Zealand. Questionnaire is the dominate data collection method in these previous studies.

Dong et al. (2017) argued that questionnaire was dominate in previous studies investigating factors influencing safety performance because there are limited data sources which contain detailed and relevant injury information. Guldenmund (2007) found that questionnaires have not been successful in uncovering factors affecting safety performance because they are more useful to expose shared attitudes rather than individual differences, which are important to predict workers' safety behavior. With more comprehensive datasets being available to construction researchers, the use of advanced data analysis techniques, e.g. machine learning, has attracted much research attention. Some data sources, e.g. the U.S. Bureau of Labor Statistics (2022), contain substantial information related to the fatalities. Natural Language Processing (NLP) has been used to extract key attributes from incident reports, based on which

a prediction model for construction safety performance can be built using machine learning (Baker et al., 2020). However, it should be noted that most studies focus on construction injury. There are limited studies which focus on fatal accidents. In construction, fatality cost is significantly higher than nonfatal injury cost. Waehrer et al. (2007) found that the average per-fatality cost is around $4m in the United States when quality of life losses is considered. On the other hand, the costliest nonfatal injuries would only cost $71,500 to $161,000. In addition, fatal accidents also have severe psychological impact on affected families and operational impact on employing organizations (Choudhry and Fang, 2008). With more comprehensive datasets and analyzing techniques being available, there is an urgent need to identify and understand factors that can influence fatal incidents in construction and build relevant predictive models.

This study therefore aims: 1) to identify indicators leading to construction fatality from a comprehensive literature review; 2) to develop predictive models of construction fatality details based on the identified explanatory indicators using machine learning approaches; and 3) to investigate the most important indicators leading to construction fatality from the predictive models.

The remaining of the paper is organized in the below structure. Section 2 provides a comprehensive literature review on factors affecting construction injury and fatality, as well as the use of machine learning approaches in construction safety-related studies. Section 3 presents the method, including data, variables, and ML approaches. Section 4 shows the results of the study and a discussion which outlines the contribution to new knowledge of this study is provided in Section 5. Section 6 concludes this study.

2. Literature review

2.1 Factors affecting construction injury and fatality

Construction injury is a complicated and multifaceted issue. Many studies have been conducted on investigating factors affecting construction injury. The first category of factors is related to construction workers' characteristics. For example, Berhanu et al. (2019) surveyed 566 construction workers in Ethiopia and found that young workers are at a higher risk of injury than workers of older age. This is probably because young workers lack comprehensive on the job training. Even when they have concerns of work conditions which may expose them to injury, they may not voice their concerns (Smith et al., 2015). In addition, male workers are more likely to be exposed to injury (Amissah et al., 2019). It is also found that the time with

employer can also affect occupational injury. Workers with longer time with employer have greater chance to occupational injury (Berhanu et al., 2019).

The second category of indicators is related to employers' characteristics, including whether written safety management plan/procedure and job training are provided, and federal standards related to safety are adopted. The establishment of an appropriate safety management program has always been considered as one of the most important strategies to reduce work injuries because it tells onsite workers exactly what they can do and prevent unsuitable behaviors (Jensen and Friche, 2007). Consequently, an effective safety management program can help reduce accident costs and improve productivity and workers' morale (Rowlinson, 2003). Job training is also found to have an impact on safety level (Rozenfeld et al., 2010). This is because most workers acquire knowledge and skills through on-the-job training (Ismail et al., 2012). Lack of strict rules, codes and standards related to safety can also contribute to safety problems (Al-Humaidi and Tan, 2010).

Another category of indicators is related to the injury and the site condition and environment when it happens. This category includes the activity that the worker is performing when injury happens, the provision of personal protection equipment (PPE), and whether appropriate site inspections have been conducted so that employers and supervisors are aware of potential hazards. Some activities are less likely to lead to injuries. For example, Amissah et al. (2019) found that steel bender/fixers are more protected from injury when compared with workers of other trade specialization. Lowery et al. (2000) found that workers who are involved in installing glass, metal or steel are at particularly high risk of injury. As for PPE, it is found that PPE can offer protection against a variety of potential hazards from physical, chemical to electrical and mechanical (Nnaji and Karakhan, 2020). Effective and thorough site inspection which can help identify safety hazards is also important and it is recommended that the effectiveness and frequency of the site inspection be reviewed at least every three months (Ho et al., 2000).

While many studies have addressed factors that affect construction safety with special focus on injury, there are limited studies targeting at construction fatality. Khodabandeh et al. (2016) analyzed 967 fatal injury reports and found that decedents' characteristics, such as age, gender, educational background, and work-related variables, such as skill training and close monitoring have significant relationship with construction fatality. In addition, it is also found that almost 60% of fatalities are caused by fall from height. This is in accordance with Assaad and El-

adaway (2021) who used data mining algorithms to investigate factors influencing construction fatality using 100 fatality reports in the United States. Five clusters of factors, including over-excavating, lack of training, lack of preventive action, lack of inspection and not following proper work procedures are identified. In these studies, some other key factors, such as whether safety training, PPE, and personal fall arrest system (PFAS) are provided to onsite workers, are still missing. It is therefore necessary to use a more comprehensive dataset that includes detailed decedent, employer, site and environment information to investigate and understand factors that can influence construction fatality.

2.2 Machine learning approaches in safety research

From a statistical point of view, previous studies in construction safety research often focus on investigating the trend in accident number and its correlation with a few circumstantial factors (Baker et al., 2020). A few recent studies have started to use machine learning to predict construction safety outcome. For example, Tixier et al. (2016) used random forest (RF) to predict injury type and body part based on 5298 injury reports from industrial, energy, infrastructure, and mining sectors. It is found that the model has high predictive capacity and outperforms traditional parametric models. In addition, Poh et al. (2018) also used RF to investigate safety leading indicators from a 7-year safety dataset of a company and found that project-related indicators, such as project type, ownership and manpower and safety-related indicators, such as lifting operations, scaffolds, platform, and safety inspection, can be used to effectively predict construction incidents. Baker et al. (2020) used more than 90,000 incident reports and machine learning to predict injury severity.

There are very limited studies which use machine learning to investigate factors leading to construction fatality and eventually build a predictive model. Choi et al. (2020) is the first study to use machine learning techniques, including logistic regression, random forest and decision tree to predict the likelihood of construction fatality. The indicators used in this study are predominately employee-related and project-related indicators, including age, length of service, type of construction, employer size, and day and month of week. Choi et al. (2020) argued that because injury-based data are not included in the model, the predictive power of the model can be limited. Another study that has a special focus on construction fatality is Mohammed and Mahmud (2020), which used machine learning to predict whether an injury is fatal or nonfatal. It is found that gradient boosted trees, random forest and decision tree can be used to effectively build the predictive models with approximate accuracy of 70%. It should be noted that this study is an exploratory study which only reported the accuracy of predictive models. Important

factors leading to fatalities were not appropriately identified. Assaad and El-adaway (2021) used spectral clustering to understand key factors causing fatality based on 100 incident reports. The sample size may limit the results and no prediction models are developed in Assaad and El-adaway (2021).

In order to further prepare for potential fatalities on construction sites, employers need more than the likelihood of fatality on site. Factors leading to fatality details, such as nature of injury, body of part, and potential exposure activities should be identified and investigated.

3. Method

3.1 Data

The National Institute for Occupational Safety and Health (NIOSH) initiated the Fatality Assessment and Control Evaluation (FACE) program in 1982. The program provides comprehensive information related to fatal injuries and is extremely useful for injury and fatality prevention. Each fatal injury case includes information collected from the employer, onsite coworkers, safety personnel, onsite emergency response team members and other witnesses. The data in this program includes 768 construction-related fatalities from 1982 to 2014 reported by FACE.

Fatal incidents in this study are coded based on the Occupational Injury and Illness Classification System (Bureau of Labor Statistics, 2021). The classification system codes the characteristics of the fatal injury on (Bureau of Labor Statistics, 2021):

1. Nature of injury or illness (NOI). It identifies the principle physical characteristics of the injury or illness. Some common types in construction fatalities include traumatic injuries to bones, nerves, and spinal cord, open wounds, burns and intracranial injuries.
2. Part of body affected (POB). It identifies the part of body directly affected by the nature of injury or illness.
3. Source and secondary source of injury or illness (SOI). It identifies the objects, substances, equipment, and other factors that lead to the injury or illness. Some common types include chemicals, machinery, structure, and vehicles.
4. Event or exposure (EOE). It describes the events where the injury or illness is produced. Some common types including contact with objects and equipment, fall, exposure to harmful substances and transportation accidents.

Table 1 shows the 4 dependent variables and measurement values.

Table 1. Dependent variables of fatal injury in this study

Dependent variables	Values
1. NOI	1. Traumatic injuries to bones, nerves 2. Open wounds 3. Burns 4. Intracranial injuries 5. Multiple traumatic injuries and disorders 6. Other traumatic injuries and disorders
2. POB	1. Head 2. Cranial region 3. Trunk 4. Lower extremities 5. Body systems 6. Multiple body parts
3. SOI	1. Chemicals 2. Containers 3. Furniture and fixtures 4. Construction, logging and mining machinery 5. Parts and materials 6. Structures and surfaces 7. Tools, instruments, and equipment 8. Vehicles 9. Temperature, environmental, debris and steam
4. EOE	1. Contact with objects and equipment 2. Fall 3. Exposure to harmful substances or environment 4. Transportation accidents 5. Fire and explosion

33 explanatory variables are originally selected to explain these four dependent variables. Table 2 summarizes the final 30 explanatory variables used in this study. 3 factors, including time with employers, fall height, and number of workers on site are excluded because many missing data are found. 13 explanatory variables (1-13) are related to actual recorded fatality details while the remaining 17 (14-30) are related to the recommendation strategies provided by FACE after fatality investigation. These 17 explanatory variables also provide critical fatality information, e.g. whether fatality is caused by non-functional PPE even though PPE may be provided by employer. A total of 694 fatalities is identified with complete data on four dependent variables and 30 explanatory variables.

Table 2. Explanatory variables of fatal injury in this study

Dependent variables	Values
Decedent characteristics	
1. Age	Years
2. Gender	Male or Female

3. Employee status	1-Wage and salary; 2-Self-employed; 3-Family business; 4-Volunteer; and 5-Not reported
Employer characteristics	
4. Onsite written safety program	1-Yes; 2-No; and 3-Not reported
5. Onsite job training provided	1-Yes; 2-No; and 3-Not reported
6. Federal standard adopted	1-Yes; 2-No; and 3-Not reported
Injury-related considerations	
7. Injury year	
8. Workers' activity	1-Vehicle related; 2-Operating equipment and machinery; 3-Tools; and 4-Construction activities.
9. Fall	1-Yes; and 2-No
Site condition and environment	
10. Location	1-Building and construction; 2-Farm; 3-Mining and other industrial areas; 4-Road; and 5-Others
11. Site inspection conducted	1-Yes; 2-No; and 3-Not reported
12. Employer was aware of hazards	1-Yes; 2-No; and 3-Not reported
13. With safety equipment	1-Yes; 2-No; and 3-Not reported
Recommended improvement strategies	
14. Provide functional PPE	1-Yes; and 2-No
15. Inspect PPE functionality	
16. Enforce proper use of PPE	
17. Provide function Personal Fall Arrest System (PFAS)	
18. Inspect PFAS functionality	
19. Enforce proper use of PFAS	
20. Provide proper equipment	
21. Inspect equipment functionality	
22. Enforce proper use of equipment	
23. Install safety protection	
24. Provide further job training	
25. Provide further safety training	
26. Further safety hazards analysis	
27. Safe worksite conditions	
28. Improve employer awareness	
29. Competent worksite safety monitoring	
30. Clear communication system	

3.2 Machine learning models

This study implements a variety of machine learning models, including Partial Least Square (PLS) (Abdi, H. 2003), Neural Network (NN) (Hecht-Nielsen, 1992), Support Vector Machines with Linear Kernel (SVMLK) (Hsu et al., 2003), Decision Tree (DT) (Safavian and Landgrebe, 1991), Random Forest (RF) (Belgiu and Drăguţ, 2016), and Stochastic Gradient Boosting Model (SGBM) (Friedman, 2002). These models are conducted using the R software

packages “caret” (Kuhn, 2008), “pls” (Wehrens and Mevik, 2007), “nnet” (Ripley et al., 2016), “kernlab” (Karatzoglou et al., 2004), “rpart” (Therneau et al., 2015), “ranger” (Wright and Ziegler, 2015), and “gbm” (Ridgeway and Ridgeway, 2004). These machine learning models, along with computing packages, are well developed and have been implemented in construction injury prediction (e.g. Baker et al., 2020). In this method section, we have provided the modelling details of decision tree because it has advantages of modelling relationships for big data sets and a large number of explanatory variables, high efficiency in computation, visualization of the tree form relationships between variables, and the form of relationship assumptions between variables is not required.

Decision tree models to predict the four fatality characteristics (NOI, POB, SOI, EOE) are developed. The decision tree -based fatal injury modelling includes the following three steps. The first step is to model relationships between each fatality characteristic and the potential explanatory variables using the decision tree models. Next, model performance is evaluated through the comparison of the six machine learning models. Finally, relative importance of variables for predicting fatality characteristics are estimated. The details of the three steps are presented in the following subsections.

3.2.1 Decision tree-based modelling of fatal injury

Decision tree is a non-parametric model of data mining for investigating nonlinear relationships between fatality characteristics and potential explanatory variables. Similar to other machine learning algorithms, it is essential to perform model validation to eliminate potential overfitting issues. The equations of decision tree-based fatal injury analysis are:

$$Y_{EOE} = f(X_1, \dots, X_{30}) \tag{1}$$

$$Y_{SOI} = f(X_1, \dots, X_{30}) \tag{2}$$

$$Y_{POB} = f(X_1, \dots, X_{30}) \tag{3}$$

$$Y_{NOI} = f(X_1, \dots, X_{30}) \tag{4}$$

where Y_{NOI} , Y_{POB} , Y_{SOI} , and Y_{EOE} are fatality characteristics, $X_1, \dots, X_{30}$ are potential explanatory variables, and f is decision tree model. In decision tree models, the explanatory variables can be both categorical and continuous variables.

It should be noted that, in reality, it is not possible to use fall (X_9) as an explanatory factor so a predictive model of likelihood of fall is also developed to overcome such limitation (see Eq. 5).

$$X_9 = f(X_1, \dots, X_8, X_{10}, \dots X_{30}) \tag{5}$$

The decision tree models are conducted using the R software package "rpart" (Therneau et al., 2015) and trained using the package "caret" (Kuhn, 2008).

3.2.2 Model validation

The accuracy of all machine learning models is evaluated using the 10-fold cross validation. In the 10-fold cross validation, data are randomly split into ten folds. Nine folds of data are used as training data for modelling, and the remained fold of data are testing data for validating predictions. The training and testing process is conducted ten times, and the cross validation accuracy of the models is the mean of the ten accuracy values. The model accuracy is evaluated using two indicators: the overall accuracy and Kappa coefficient. The equations of the two indicators are:

$$O = \frac{TP+TN}{T} \tag{7}$$

$$\kappa = \frac{T(TP+TN)-\Sigma}{T^2-\Sigma} \tag{8}$$

where TP is the true positive classification, i.e., the number of correctly predicted data in a given class, TN is the true negative classification, i.e., the number of correctly rejected data in other classes, T is the total number of correctly predicted classes, and Σ is the change accuracy computed as $\Sigma = (\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN}) + (\mathrm{FN} + \mathrm{TN})(\mathrm{FP} + \mathrm{TN})$. The false positive (FP) and false negative (FN) classifications are the number of incorrectly predicted data in a given class, and the number of incorrectly predicted data in other classes, respectively.

3.2.3 Importance of explanatory variables

The decision tree provides a relative importance value for each explanatory variable. In decision tree, the misclassification error is calculated as the sum of squared errors (SSE) when explanatory variables split data into two regions R_1 and R_2:

$$SSE = \sum_{i \in R_1} \left(y_i - \bar{d}_1\right)^2 + \sum_{i \in R_2} \left(y_i - \bar{d}_2\right)^2 \tag{6}$$

where y_i is an observation of fatal injury, and $\bar{d}_1$ and $\bar{d}_2$ are mean values of observations in the two regions, respectively. The aim of decision tree is to find the split nodes that can minimize SSE values. The variable importance is measured by evaluating the sum of decrease in SSE when splitting an explanatory variable in the cross validation models of decision tree. The variable with the highest impact on the variation of SSE has the highest value of relative importance. The relative importance of other variables is relative decreased in SSE compared with the variable with the highest importance. In this study, to adequately compare explanatory variables, the relative importance is normalized to values that their sum is 1 through the following equation:

$$\Delta_j = \delta_j / \sum \delta_j \quad (7)$$

where Δ_j is the normalized relative importance of the variable X_j, and δ_j is the relative importance calculated with the decrease in SSE.

4. Results

4.1 DT models of fatality characteristics

Figure 1 demonstrates the statistical summary of explanatory variables within fatal event or exposure for explaining fatal event or exposure. Fatal event or exposure was identified using six explanatory variables, including fall (X_9), injury year (X_7), workers' activity (X_8), location (X_{10}), lack of properly enforced PPE use (X_{22}), and safety equipment (X_{13}). Figure 1 also shows that fall was the primary variable that controlled the fatal event or exposure, hence leading to a predictive modelling being developed for the likelihood of fall in this study. According to the variables in the top three layers, the fatal event or exposure was categorized into five groups: the first group contains fall (X_9=1), the second group contains contact with objects and equipment and transportation accidents (X_9=2; $X_7 \geq 1995$; X_8=2,3,4), the third group contains transportation accidents (X_9=2; $X_7 \geq 1995$; X_8=1), the fourth group contains exposure to harmful substances or environment (X_9=2; X_7<1995; X_8=2,3,4) and the fifth group of transportation activities (X_9=2; X_7<1995; X_8=1).

Similarly, the statistical summary of explanatory variables within other fatality characteristics and fall are provided in Appendix 1.

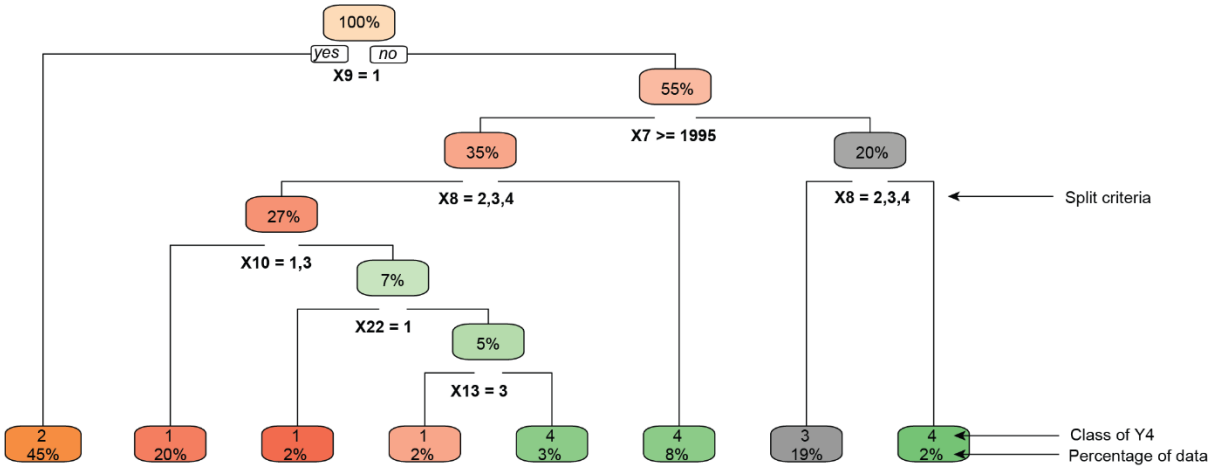


Figure 1. Statistical summaries of explanatory variables within fatal event or exposure for explaining fatal event or exposure

4.2 Model validation

Table 3 shows the prediction accuracy of all six models and relevant Kappa coefficient when predicting the four fatality characteristics.

It is found that of the six models investigated in this study, decision tree, random forest and stochastic gradient boosting model perform relatively well against the other three. For example, SGBM has a prediction accuracy of 84.7% of construction fall with a Kappa coefficient of 0.689. Kappa coefficient represents the consistency of classification and identification by fair (0.21-0.40), moderate (0.41-0.60), and substantial (0.61-0.80) and almost perfect (0.81-1.00) (Li et al., 2020).

Table 3. Accuracy and Kappa coefficient of six ML models in predicting fatality characteristics

Fatality characteristics (Dependent ~ explanatory)	Accuracy						Kappa					
	PLS	NN	SVMLK	DT	RF	SGBM	PLS	NN	SVMLK	DT	RF	SGBM
NOI ~ x1-x30	0.510	0.512	0.533	**0.566**	0.562	0.565	0.249	0.269	0.315	0.351	0.347	**0.352**
POB ~ x1-x30	0.461	0.387	0.425	0.496	0.532	**0.540**	0.208	0.154	0.243	0.250	0.330	**0.347**
SOI ~ x1-x30	0.647	0.545	0.571	0.690	**0.765**	0.732	0.459	0.376	0.434	0.549	**0.658**	0.609
EOE ~ x1-x30	0.712	0.772	0.843	0.784	**0.849**	0.847	0.571	0.674	0.775	0.689	**0.784**	0.781
NOI ~ Y3,Y4,x1-x30	0.513	0.523	0.565	0.519	0.558	**0.569**	0.254	0.286	0.363	0.272	0.343	**0.363**
POB ~ Y3,Y4,x1-x30	0.502	0.446	0.477	0.559	0.562	**0.586**	0.264	0.245	0.322	0.366	0.374	**0.404**
Fall ~ x1-x8,x10-x30	0.793	0.817	0.829	0.801	0.823	**0.847**	0.578	0.627	0.650	0.592	0.638	**0.689**

In addition, when predicting the four fatality characteristics, it appears that the three models can be used to predict fatality sources and events (SOI and EOE) more effectively. The prediction accuracy of SOI and EOE is 0.765 and 0.849 respectively using RF and relevant Kappa coefficients are 0.658 and 0.784, indicating substantial consistency. As for predicting fatality types (NOI and POB), the accuracy is relatively lower at 0.566 (using DT) and 0.540 (using SGBM), with Kappa coefficients of 0.352 and 0.347, indicating fair consistency.

There may be a causal relationship between potentially fatal events and activities (i.e. SOI and EOE) and fatality type (i.e. NOI and POB), so additional two predictive models of $Y_{NOI} = f(Y_3, Y_4, X_1, \dots, X_{30})$ and $Y_{POB} = f(Y_3, Y_4, X_1, \dots, X_{30})$ are also tested to examine whether prediction accuracy can be improved by integrating exposure or event and source of injury in the prediction. The results show that there are very small differences by including a phased prediction model (i.e. to predict NOI and POB with predicted SOI and EOE), even though a causal relationship exists. The prediction accuracy of NOI and POB is improved from 0.566 to 0.569 and 0.540 to 0.586, with very slight and negligible difference.

4.3 Importance of explanatory variables

Figure 2 shows the relative importance of explanatory variables when predicting NOI. The most important indicators are fall (X_9, 0.36), functional PFAS (X_{17}, 0.15), injury year (X_7, 0.12), proper use of PFAS (X_{19}, 0.12), safety equipment (X_{13}, 0.08), location (X_{10}, 0.07), install safety protection (X_{23}, 0.04) and workers' activity (X_8, 0.03).

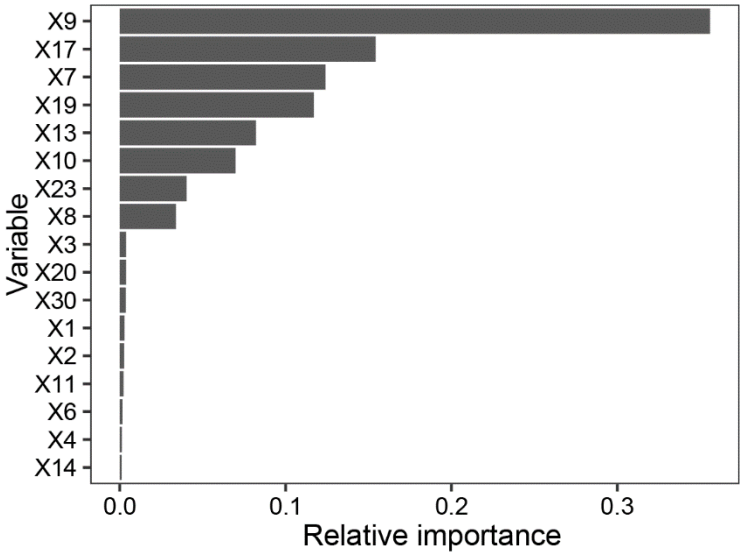


Figure 2. Relative importance of explanatory variables for predicting NOI

Figure 3 shows the relative importance of explanatory variables when predicting POB. The most important indicators are injury year (X_7, 0.31), fall (X_9, 0.24), functional PFAS (X_{17}, 0.11), proper use of PFAS (X_{19}, 0.09), workers' activity (X_8, 0.08), safety equipment (X_{13}, 0.06), and location (X_{10}, 0.04).

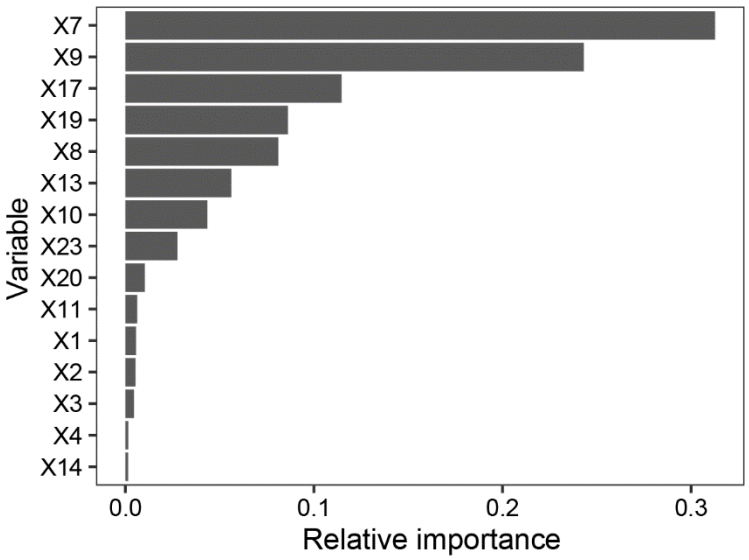


Figure 3. Relative importance of explanatory variables for predicting POB

Figure 4 shows the relative importance of explanatory variables when predicting SOI. The most important indicators are fall (X_9, 0.36), functional PFAS (X_{17}, 0.16), proper use of PFAS (X_{19}, 0.12), safety equipment (X_{13}, 0.08), injury year (X_7, 0.07), workers' activity (X_8, 0.06), and install safety protection (X_{23}, 0.04)

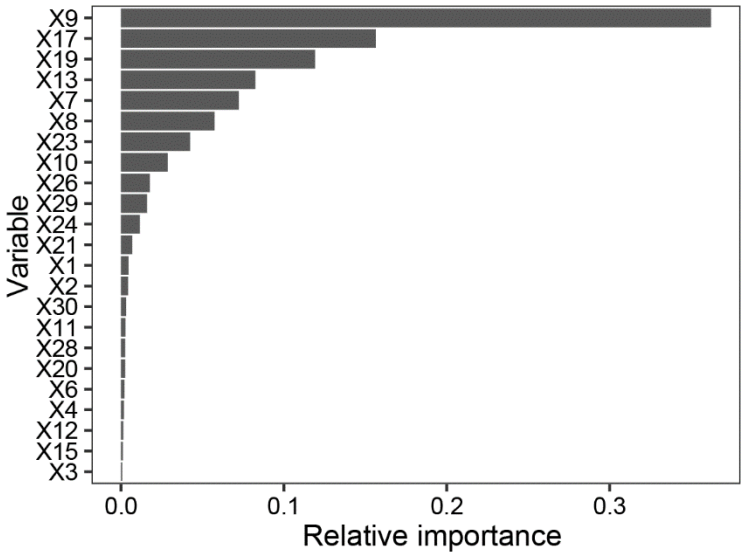


Figure 4. Relative importance of explanatory variables for predicting SOI

Figure 5 shows the relative importance of explanatory variables when predicting EOE. The most important indicators are fall (X_9, 0.37), functional PFAS (X_{17}, 0.16), proper use of PFAS (X_{19}, 0.12), workers' activity (X_8, 0.09), safety equipment (X_{13}, 0.09), injury year (X_7, 0.08), and install safety protection (X_{23}, 0.04).

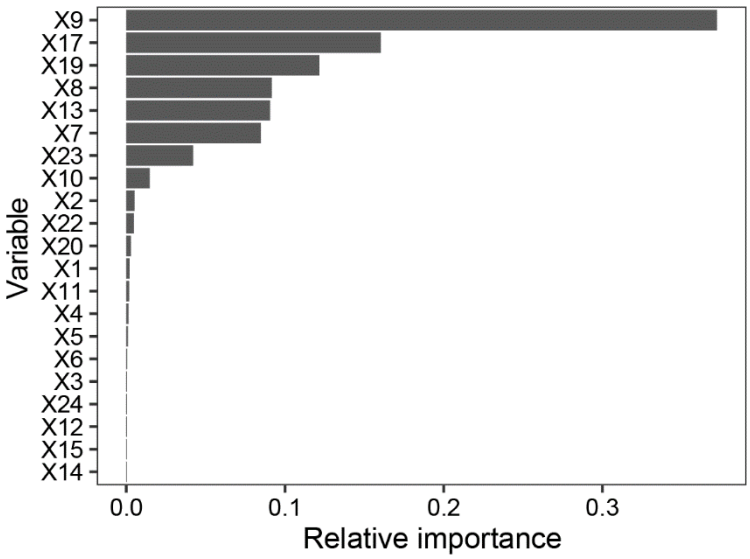


Figure 5. Relative importance of explanatory variables for predicting EOE

The results show that the risk of fall is a critical indicator for predicting fatality source and type. The predictive model developed in this study can also help predict the likelihood of fall with an accuracy of 84.7%. As can be seen from Figure 6, the most important explanatory variables include functional PFAS (X_{17}, 0.43), proper use of PFAS (X_{19}, 0.23), location (X_{10}, 0.06), workers' activity (X_8, 0.06), safety equipment (X_{13}, 0.05) and safety worksite condition (X_{27}, 0.05). This predictive model will help determine a likelihood of fall that can be used in the predictive models of fatality characteristics.

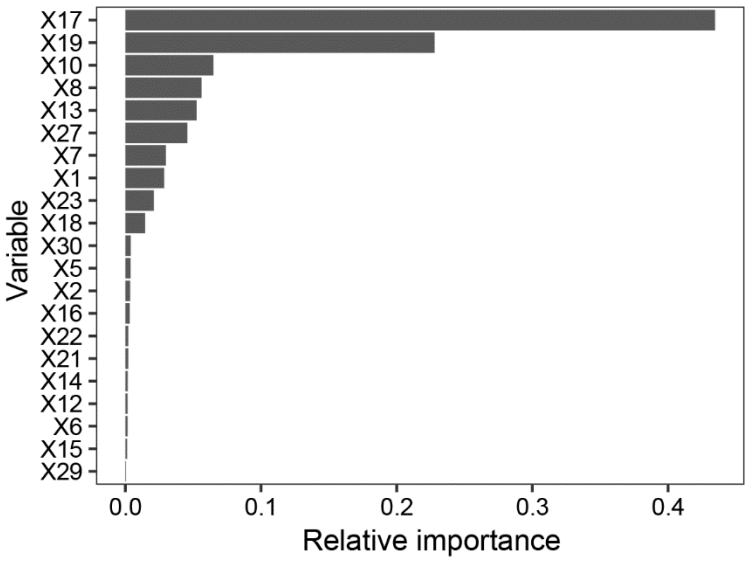


Figure 6. Relative importance of explanatory variables for predicting fall

5. Discussion

The contribution to new knowledge of this study is at three levels. This is one of the first studies to comprehensively investigate factors leading to construction fatality using machine learning and eventually predictive models are developed to predict fatality characteristics. While using machine learning to predict construction injury has been addressed in many recent studies (e.g. Tixier et al., 2016; Poh et al., 2018; Baker et al., 2020), construction fatality has only been investigated in three recent studies, which are Choi et al. (2020), Mohammed and Mahmud (2020) and Assaad and El-adaway (2021). Given the heavy financial and emotional implications of construction fatalities, exploring and understanding key factors and eventually developing predictive models are extremely important.

Instead of developing only a predictive model for the likelihood of construction fatality, which has been partially addressed in Choi et al. (2020), this study also develops predictive models of fatality characteristics, including nature of injury, part of body, source of injury and event or exposure. Knowing these potential fatality related characteristics and understanding key factors influencing fatality characteristics will better prepare employers and onsite management team for establishing appropriate safety plans and emergency response plans. The results show that compared with Choi et al. (2020), even though many injury-related explanatory variables, such as workers' activity, location, completed site inspection, on-the-job training, PFAS and its functionality, have now been included, ML models can still provide a relatively accurate prediction of the likelihood of fall, with a prediction accuracy of 84.7%

using SGBM. Given the importance of fall when predicting fatality characteristics, including NOI, POB, SOI and EOE, an accurate prediction of the likelihood of fall is the first important step. The results show that the most important indicators to predict the likelihood of fall are PFAS (including functionality of the PFAS and whether employees are appropriately trained on the use of PFAS), location and workers' activity, safety equipment (e.g. hard hat, work gloves and boots) and safety work conditions. This ML-based model and findings are new and have not been addressed in previous studies.

In addition, the four predictive models of NOI, POB, SOI and EOE also provide some useful insights. The predictive models have relatively higher accuracy when predicting source of injury (76.5%) and fatal exposure or event (84.9%). This is important to ensure that appropriate safety plans and emergency response plans are available to address potentially fatal events. For example, construction sites that are subject to the risk of fall can use harness with embedded Bluetooth low energy (BLE) devices (Rey-Merchán et al., 2022). On the other hand, construction sites that are more prone to fatality caused by contact with objects and equipment could focus on machine guarding and proximity warning (Perlman et al., 2014). Important indicators for predicting SOI and EOE are largely the same, with the most important ones being the likelihood of fall, PFAS (functionality and relevant training), workers' activity, onsite safety equipment and install safety protection.

The predictive models for nature of injury and part of body are also developed. Nature of injury and part of body are directly related to first aid facilities and services that should be available on site. Anantharaman et al. (2022) found that casualties of construction incidents in Singapore are not provided with any on-site first aid immediately after the incident. In case of severe injuries, first aid services will only be provided when visiting a clinic or hospital emergency department. It is concluded that lack of prevision of first aid services on construction site can delay emergency care and should be rectified immediately. The prediction accuracy of NOI and POB is relatively lower at 0.566 and 0.540. However, they still offer a fair prediction power for determining appropriate first aid services when compared with previous models on injury prediction (see Baker et al., 2020).

A comparison of the prediction accuracy with previous studies using ML in safety research has also been conducted. Kang and Ryu (2019) used random forest to predict types of occupational accidents at construction sites at an accuracy of 71.3%. The RF model in Choi et al. (2020) to predict the likelihood of construction fatality is 91.98%. However, it should be noted that this

RF model only includes limited employee-related and project-related indicators and no injury-related data are included, explaining why the accuracy score is relatively higher. Baker et al. (2020) compared the performance of random forest, extreme gradient boosting and linear support vector machine and found that the three models perform comparably. For example, the prediction powers of the three models are 60%, 63% and 65% for injury severity, 55%, 55% and 56% for injury type, 42%, 42% and 42% for incident type and 38%, 39% and 37% for body part. Our findings suggest that decision tree, random forest and stochastic gradient boosting model perform relatively well in predicting fall and the four fatality characteristics. The models developed in this study have distinct advantages in the number of explanatory variables included and prediction accuracy when compared with previous models. Specifically, when comparing with Baker et al. (2020), the prediction accuracy of POB is increased from near 40% to 54% and the prediction accuracy of incident type (e.g. SOI and EOE) is increased from near 40% to 76%-85%, depending on the ML models.

6. Conclusions

This study analyzed 694 fatalities reported in the Fatality Assessment and Control Evaluation (FACE) program from the National Institute for Occupational Safety and Health (NIOSH) to develop predictive models of fatality characteristics, including nature of injury, part of body, source of injury and event or exposure. The study makes significant contribution to new knowledge because it is one of the first studies to include a total of 30 explanatory variables that comprehensive represent employee, employer, project, site and fatality-related information when developing the predictive models. The explanation power of the 30 explanatory variables is strong, as evidenced by the prediction accuracy of 56.6%, 54.0%, 76.5% and 84.9% for NOI, POB, SOI, EOE respectively. When predicting the likelihood of fall, a major fatal exposure or event in construction, the accuracy is 84.7% with substantial consistency observed. It is also found that decision tree, random forest and stochastic gradient boosting model have comparable performance when predicting fatality characteristics. When compared with partial least square, neural network and support vector machines with linear kernel, the prediction accuracy is consistently higher.

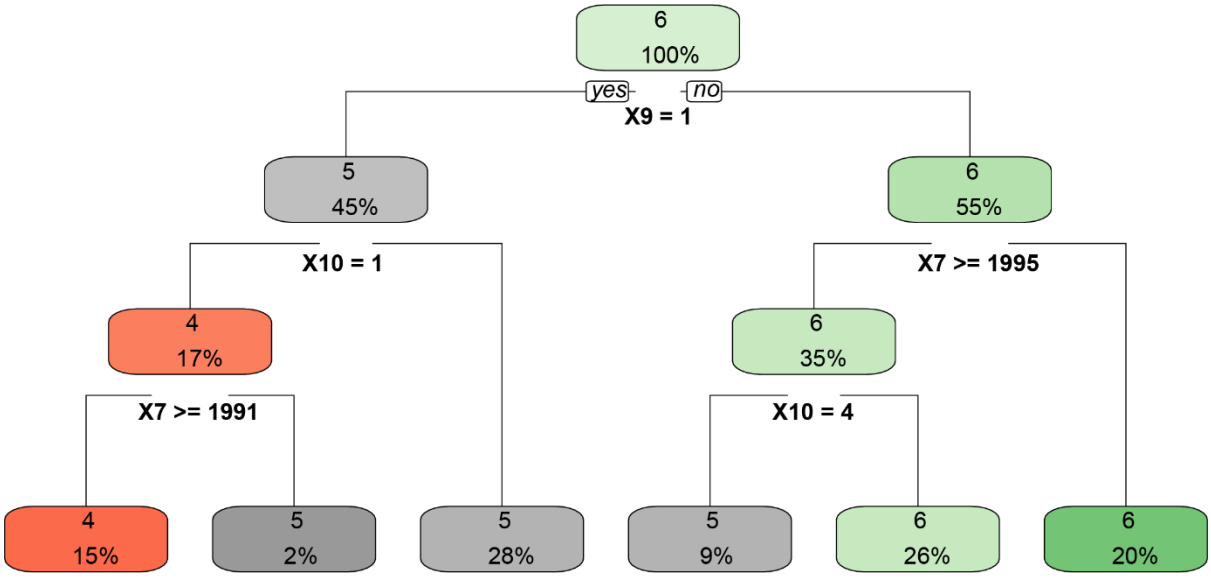
In addition, this study also identifies important indicators when predicting fatality characteristics. SOI and EOE are mostly affected by the likelihood of fall, PFAS (functionality and relevant training), workers' activity, onsite safety equipment and install safety protection. NOI and POB are mostly affected by the likelihood of fall, PFAS, injury year, workers' activity,

location and safety equipment. Understanding these important indicators and eventually using them to predict potential fatality types, body parts, and exposure activities and events are extremely important to set up onsite emergency response plans and first aid services and facilities.

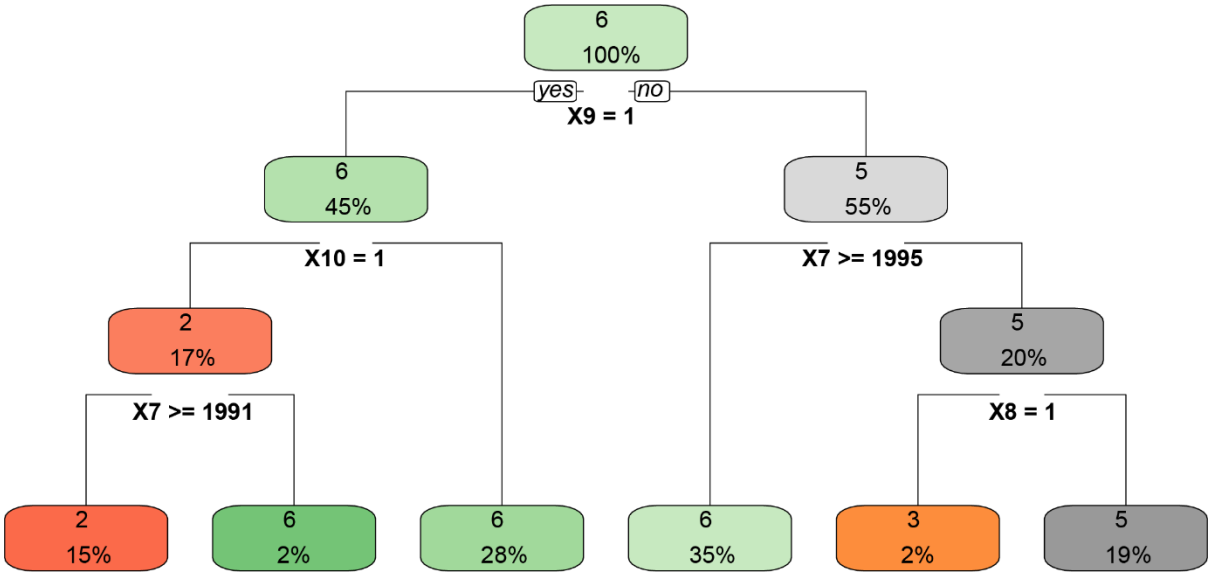
This study has a few limitations. The dataset is related to the U.S. construction industry and may not be directly used by other countries, although similar process and method can be adopted if comparable datasets can be obtained from other countries. In addition, some factors, e.g. time with employers and number of workers on site, are excluded due to lack of complete data. The inclusion of additional explanatory factors may further enhance the prediction power of the ML models in this study.

Appendix 1

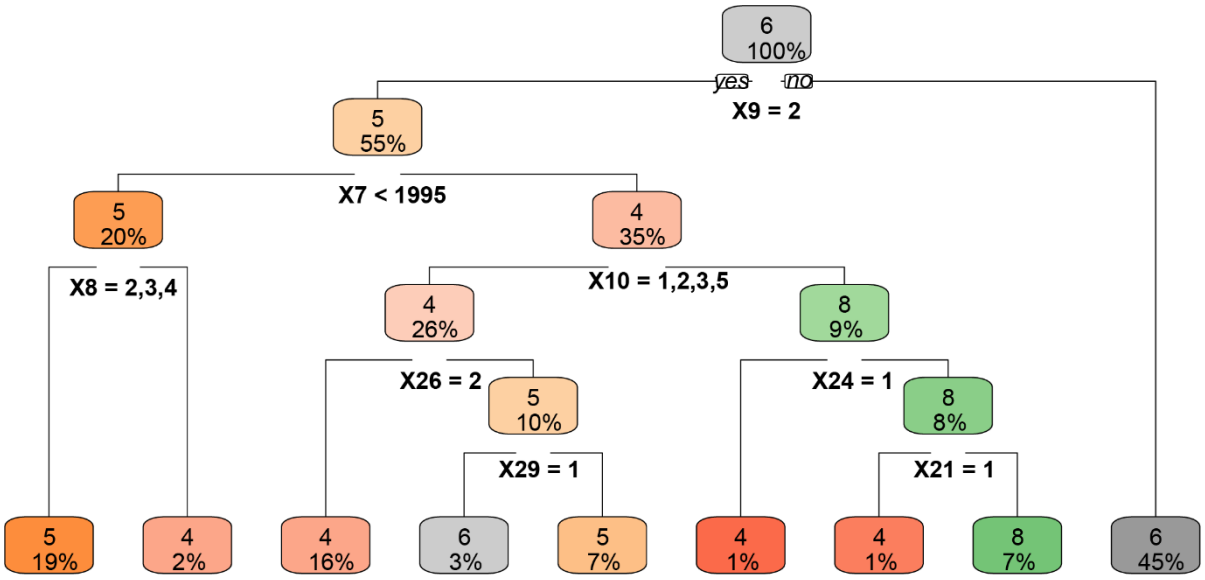
The statistical summary of explanatory variables within fatality characteristics and fall



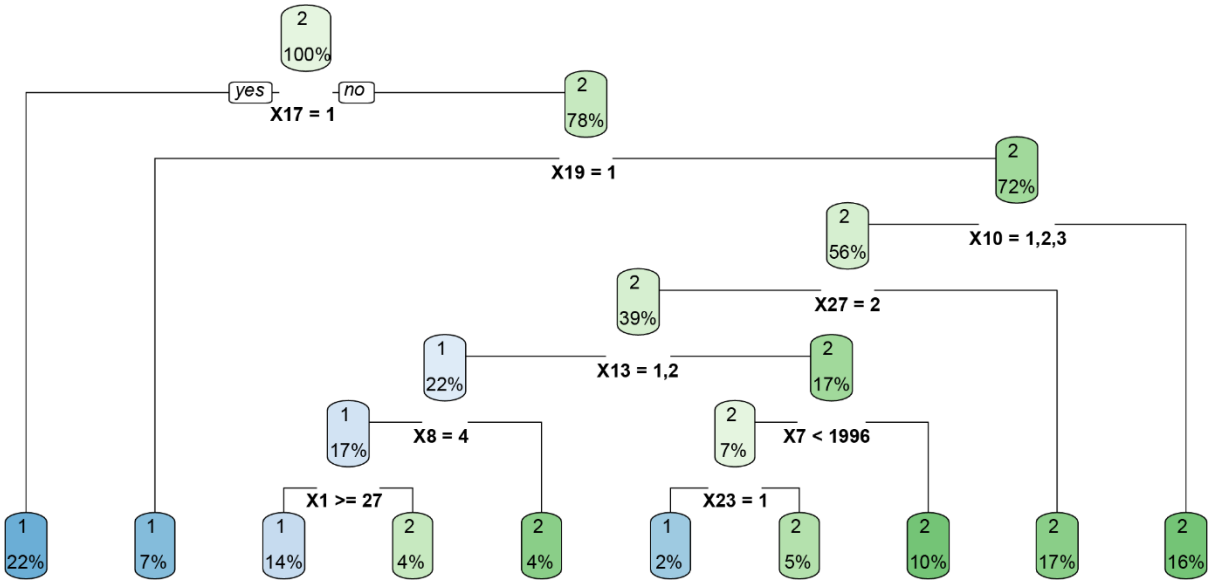
Appendix 1a. Statistical summaries of explanatory variables within NOI



Appendix 1b: Statistical summaries of explanatory variables within POB



Appendix 1c: Statistical summaries of explanatory variables within SOI



Appendix 1d: Statistical summaries of explanatory variables within fall

References

Abdi, H. (2003). Partial least square regression (PLS regression). Encyclopedia for research methods for the social sciences, 6(4), 792-795.

Al-Humaidi, H. M., & Tan, F. H. (2010). Construction safety in Kuwait. Journal of Performance of Constructed Facilities, 24(1), 70-77.

Amissah, J., Badu, E., Agyei-Baffour, P., Nakua, E. K., & Mensah, I. (2019). Predisposing factors influencing occupational injury among frontline building construction workers in Ghana. BMC research notes, 12(1), 1-8.

Anantharaman, V., Zuhary, T. M., Ying, H., & Krishnamurthy, N. (2022). Characteristics of injuries resulting from falls from height in the construction industry. Singapore medical journal. https://doi.org/10.11622/smedj.2022017

Assaad, R., & El-adaway, I. H. (2021). Determining critical combinations of safety fatality causes using spectral clustering and computational data mining algorithms. Journal of Construction Engineering and Management, 147(5), 04021035.

Baker, H., Hallowell, M. R., & Tixier, A. J. P. (2020). AI-based prediction of independent construction safety outcomes from universal attributes. Automation in Construction, 118, 103146.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 114, 24-31.

Berhanu, F., Gebrehiwot, M., & Gizaw, Z. (2019). Workplace injury and associated factors among construction workers in Gondar town, Northwest Ethiopia. BMC musculoskeletal disorders, 20(1), 1-9.

Bureau of Labor Statistics. (2021). Occupational injury and illness classification manual. Available at: https://www.bls.gov/iif/oshoiics.htm#:~:text=The%20Occupational%20Injury%20and%20Illness,Fatal%20Occupational%20Injuries%20(CFOI) (Cited 01 May 2022)

Choi, J., Gu, B., Chin, S., & Lee, J. S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. Automation in Construction, 110, 102974.

Choudhry, R. M., & Fang, D. (2008). Why operatives engage in unsafe work behavior: Investigating factors on construction sites. Safety science, 46(4), 566-584.

Dong, X. S., Largay, J. A., Wang, X., Cain, C. T., & Romano, N. (2017). The construction FACE database—Codifying the NIOSH FACE reports. Journal of safety research, 62, 217-225.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

Guldenmund, F. W. (2007). The use of questionnaires in safety culture research–an evaluation. Safety science, 45(6), 723-743.

Guo, B. H., Yiu, T. W., & González, V. A. (2016). Predicting safety behavior in the construction industry: Development and test of an integrative model. Safety science, 84, 1-11.

Health and Safety Executive. (2022). Work-related fatal injuries in Great Britain. Available at: https://www.hse.gov.uk/statistics/fatals.htm (Cited 26 Apr 2022)

Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In Neural networks for perception (pp. 65-93). Academic Press.

Ho, D. C. P., Ahmed, S. M., Kwan, J. C., & Ming, F. Y. W. (2000). Site safety management in Hong Kong. Journal of Management in Engineering, 16(6), 34-42.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

Ismail, Z., Doostdar, S., & Harun, Z. (2012). Factors influencing the implementation of a safety management system for construction sites. Safety science, 50(3), 418-423.

Jensen, L. K., & Friche, C. (2007). Effects of training to implement new tools and working methods to reduce knee load in floor layers. Applied ergonomics, 38(5), 655-665.

Kang, K., & Ryu, H. (2019). Predicting types of occupational accidents at construction sites in Korea using random forest model. Safety Science, 120, 226-236.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab-an S4 package for kernel methods in R. Journal of statistical software, 11(9), 1-20.

Khodabandeh, F., Kabir-Mokamelkhah, E., & Kahani, M. (2016). Factors associated with the severity of fatal accidents in construction workers. Medical journal of the Islamic Republic of Iran, 30, 469.

Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of statistical software, 28, 1-26.

Li, Y., Wang, J., Gao, T., Sun, Q., Zhang, L., & Tang, M. (2020). Adoption of machine learning in intelligent terrain classification of hyperspectral remote sensing images. Computational Intelligence and Neuroscience, 2020.

Lowery, J. T., Glazner, J., Borgerding, J. A., Bondy, J., Lezotte, D. C., & Kreiss, K. (2000). Analysis of construction injury burden by type of work. American journal of industrial medicine, 37(4), 390-399.

Mohammed, J., & Mahmud, M. J. (2020). Selection of a machine learning algorithm for OSHA fatalities. In 2020 IEEE Technology & Engineering Management Conference (TEMSCON) (pp. 1-5). IEEE.

Nnaji, C., & Karakhan, A. A. (2020). Technologies for safety and health management in construction: Current use, implementation benefits and limitations, and adoption barriers. Journal of Building Engineering, 29, 101212.

Perlman, A., Sacks, R., & Barak, R. (2014). Hazard recognition and risk perception in construction. Safety science, 64, 22-31.

Poh, C. Q., Ubeynarayana, C. U., & Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach. Automation in construction, 93, 375-386.

Rey-Merchán, M. D. C., Gómez-de-Gabriel, J. M., Fernández-Madrigal, J. A., & López-Arquillos, A. (2022). Improving the prevention of fall from height on construction sites through the combination of technologies. International journal of occupational safety and ergonomics, 28(1), 590-599.

Ridgeway, G., & Ridgeway, M. G. (2004). The gbm package. R Foundation for Statistical Computing, Vienna, Austria, 5(3).

Ripley, B., Venables, W., & Ripley, M. B. (2016). Package 'nnet'. R package version, 7(3-12), 700.

Rowlinson S.M. (2003). Hong Kong Construction: Safety Management and the Law. Sweet & Maxwell Asia Causeway Bay; Hong Kong, China.

Rozenfeld, O., Sacks, R., Rosenfeld, Y., & Baum, H. (2010). Construction job safety analysis. Safety science, 48(4), 491-498.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.

Smith, P. M., Saunders, R., Lifshen, M., Black, O., Lay, M., Breslin, F. C., ... & Tompa, E. (2015). The development of a conceptual model and self-reported measure of occupational health and safety vulnerability. Accident Analysis & Prevention, 82, 234-243.

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2015). Package 'rpart'. Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016).

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Application of machine learning to construction injury prediction. Automation in construction, 69, 102-114.

U.S. Bureau of Labor Statistics. (2022). Census of fatal occupational injuries (CFOI). Available at: https://www.bls.gov/iif/oshcfoi1.htm (Cited 26 Apr 2022)

U.S. Department of Labor. (2022). Occupational Safety and Health Administration: commonly used statistics. Available at: https://www.osha.gov/data/commonstats (cited 26 Apr 2022)

Xia, N., Zou, P. X., Liu, X., Wang, X., & Zhu, R. (2018). A hybrid BN-HFACS model for predicting safety performance in construction projects. Safety science, 101, 332-343.

Waehrer, G. M., Dong, X. S., Miller, T., Haile, E., & Men, Y. (2007). Costs of occupational injuries in construction in the United States. Accident Analysis & Prevention, 39(6), 1258-1266.

Wehrens, R., & Mevik, B. H. (2007). The pls package: principal component and partial least squares regression in R.

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.

Yu, Q. Z., Ding, L. Y., Zhou, C., & Luo, H. B. (2014). Analysis of factors influencing safety management for metro construction in China. Accident Analysis & Prevention, 68, 131-138.