ARTICLE IN PRESS

Automation in Construction xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

# Safety leading indicators for construction sites: A machine learning approach

Clive Q.X. Poh[a], Chalani Udhyami Ubeynarayana[b], Yang Miang Goh[c,*]

[a] Dept. of Building, School of Design and Environment, National Univ. of Singapore, 4 Architecture Dr., Singapore 117566, Singapore
[b] Neurodevelopment Research Centre, Singapore Institute for Clinical Sciences (SICS); Formerly Safety and Resilience Research Unit (SaRRU), Dept. of Building, School of Design and Environment, National Univ. of Singapore, Singapore
[c] Safety and Resilience Research Unit (SaRRU), Dept. of Building, School of Design and Environment, National Univ. of Singapore, 4 Architecture Dr., Singapore 117566, Singapore

ARTICLE INFO

*Keywords:*
Machine learning
Construction safety
Leading indicators
Accident prevention
Safety management

ABSTRACT

The construction industry is one of the most dangerous industries in many countries. To improve the situation, senior managers overseeing portfolios of construction projects need to understand the safety risk levels of their projects so that interventions can be implemented proactively. Safety leading indicators is one way to flag sites that are of higher risk. However, there is a lack of validated leading indicators that can reliably classify sites according to their safety risk levels. On the other hand, despite the success of machine learning (ML) approaches in other domains, it is not widely utilized in the construction industry, especially in the development of safety leading indicators. This paper presents a ML approach to developing leading indicators that classify sites in accordance to their safety risk in construction projects. This study was guided by the industry-recognized Cross-Industry Standard Process for Data Mining (CRISP-DM) framework and the key types of data used include safety inspection records, accident cases and project-related data. These data were obtained from a large contractor in Singapore and the data were accumulated from year 2010 to 2016. Out of thirty-three input variables (also known as features or independent variables), 13 input variables were selected using a combination of Boruta feature selection technique and decision tree. Of the 13 selected input variables, six of them are project-related (project type, project ownership, contract sum, percent completed, magnitude of delay and project manpower) and seven of them are items in the contractor's safety inspection checklists (crane/lifting operations, scaffold, mechanical-elevated working platform, falling hazards/openings, environmental management, good practices and weighted safety inspection score). Five popular ML algorithms were then used to train models for prediction of accident occurrence and severity. During validation, random forest (RF) provided the best prediction performance with an accuracy of 0.78 and has achieved a substantial strength of agreement with Weighted-Kappa Statistics of 0.70. Comparing with similar studies, this result is promising. The prediction (i.e. the output variable) provided by the RF model can be used as a safety leading indicator of the risk level of a site. It is recommended that the predictive RF model be deployed in construction organizations, especially large public and private developers, contractors and industry associations, to provide monthly forecast of project safety performance so that pre-emptive inspections and interventions can be implemented in a more targeted manner.

## 1. Introduction

According to the International Labor Organization [1], there is about 318,000 work-related fatalities each year and the construction industry contributes to a significant portion of the fatalities. For example, from 2012 to 2014, there were 1932 construction-related fatalities in China [46]. In 2015, there were 952 construction-related fatalities in United States [6]. In 2009, the number of fatalities in the construction industry in UK was the highest compared to other industry. Similarly, in 2012, the South Korean construction industry had the highest number of fatalities as compared to other industries [76]. Such phenomenon is also observed in Singapore, where its construction industry remains the top contributor of workplace fatalities with 24 deaths in 2016 [67]. To improve the situation, senior managers, e.g. Director of Projects, Chief Risk Officers and Corporate Workplace Safety and Health (WSH) Manager, overseeing portfolios of projects need to understand the safety risk levels of their projects so that interventions can be implemented proactively. Research has shown

* Corresponding author.
*E-mail address:* bdggym@nus.edu.sg (Y.M. Goh).

https://doi.org/10.1016/j.autcon.2018.03.022
Received 11 October 2017; Received in revised form 9 March 2018; Accepted 13 March 2018
0926-5805/ 

Please cite this article as: Poh, C.Q.X., Automation in Construction (2018), https://doi.org/10.1016/j.autcon.2018.03.022

**Table 1**
Machine learning algorithms applied in the construction project management research.

| Construction project management area | Machine learning algorithms | | | |
|---|---|---|---|---|
| | Classification (n = 5) | Regression (n = 13) | Association (n = 4) | Clustering (n = 1) |
| Cost | – | • Neuro-fuzzy [70]<br>• Generic algorithm, fuzzy logic, neural networks [13]<br>• Neuro-fuzzy [74]<br>• Principle component analysis, support vector machine [60]<br>• General regression, neural networks [3]<br>• Support vector regression, differential evolution [12]<br>• Singular value decomposition, Ridor rules, k-star, neural networks [65]<br>• Principle component analysis, support vector machine [75] | – | – |
| Safety | • Neural networks [24]<br>• Neural networks, decision trees [23]<br>• Random forest, stochastic gradient tree boosting [62] | – | • Apriori [11,47]<br>• Classification and regression tree [10] | – |
| Dispute | • Support vector machine, neural networks, decision trees [15] | – | – | – |
| Information system | – | • Neural networks, decision trees [38]<br>• Neuro-fuzzy [73]<br>• Case based reasoning, neural networks, neuro-fuzzy [71] | • Fuzzy logic, neural networks, generic algorithms [72] | • K-means clustering [2] |
| Scheduling | • Decision trees, neural networks [39] | • EM clustering [32] | – | – |
| Procurement | – | • AutoRegression tree [18] | – | – |

that senior managers who are not in the know of safety risks cannot display effective safety leadership [26,56], and ineffective safety leadership can easily lead to poor safety culture and accidents [78,79].

Large construction organizations like public developers and multinational contractors can have hundreds or even thousands of live projects at any time. However, senior managers typically lack time to plough through a wide variety of numerical data and volumes of project documents and information submitted by the different projects. Moreover, they may not have the expertise to understand all the detailed information available to them. At the same time, the number of staff assisting senior management overseeing these large number of projects may be limited. To optimize the use of limited human resources, it is important to focus the attention of available personnel on projects with higher safety risk. One way to facilitate this is to develop safety indicators to help senior managers identify projects with higher safety risk.

According to Robson et al. [58] there are two main types of safety indicators: leading and lagging indicators. Lagging indicators (also known as trailing, reactive or negative indicators) measure workplace safety and health outcomes such as injury and illness rates. On the other hand, leading indicators (also known as proactive or positive indicators) measure "workplace activities, conditions, and events" that are relevant to or may determine workplace safety and health outcomes. Examples of leading indicators include aggregated training effectiveness score, safety climate measures, and number of inspections. Safety leading indicators function as predictive metrics of safety performance that facilitates monitoring and proactive interventions to prevent accidents and ill health in the workplace [5,27]. Due to the reactive and delayed nature of lagging indicators, managers need to develop suitable leading indicators to help them assess the safety and health risk of their workplace. Safety leading indicator is also closely related to predictive models of accident which are used to forecast safety and health performance. However, past research on safety leading indicators were frequently conceptual or theoretical in nature (e.g. [57]), while studies on predictive models of accident were frequently limited by statistical assumptions (e.g. [16]) or were reliant on audit data that are not timely enough for prompt actions by management (e.g. [24]). In short, validated safety leading indicators should be developed to help organizations prioritize their effort and resources effectively [5].

Evident by its numerous applications across different fields and disciplines, machine learning (ML), a subset of artificial intelligence, have been lauded to be an effective predictive tool [40,66]. To this end, ML has in many cases matched or surpassed human experts in performing predictive tasks, e.g. load forecasting and diagnostics [66]. ML had also been shown to be able to predict non-linear and complex phenomenon (e.g. construction accidents [25]), and help management make better decisions. Despite the numerous safety leading indicators studies, research on use of ML to develop predictive leading indicators is scarce [51] and more of such research is needed [57]. Therefore, this paper aims to use a ML approach to develop a predictive model of accident occurrence and severity, i.e. a ML model capable of providing a validated safety leading indicator, to help construction organizations forecast project safety risk. It is believed that validated leading indicators will enable effective safety leadership and hence prevent accidents [26].

## 2. Literature review

### 2.1. ML studies in construction

ML can enhance management's decision making process during long term planning and day-to-day operations. According to the literature, ML had contributed towards construction project management areas such as time [39], cost [65], quality [49], safety [23] and operational performance [68]. Prediction of construction project performance gained significant interests. Some examples include prediction of cost performance of commercial building projects [60], dispute propensity in public-private partnership projects [15] and accident occurrence and severity [24]. ML often performed better than existing methods in situations where the domain problem is premised upon a confluence of factors. For instance, Hammad et al. [32] have reported that the estimation of the duration of steel fabrication using ML has outperformed existing estimating technique. In another study, Yu & Skibniewski [74]