https://dx.doi.org/10.1016/j.jclepro.2020.122266

Detecting Corporate Misconduct through Random Forest in China's Construction Industry

Ran Wang[1], Vahid Asghari[2], Shu-Chien Hsu[3], Chia-Jung Lee[4], and Jieh-Haur Chen[5]

ABSTRACT

Construction companies' wrongdoings can result in severe consequences and have been a concern for regulators, investors, and other stakeholders. Though previous studies have identified a great number of factors associated with corporate misconduct, ranking their importance and using them to predict this misconduct in the construction industry have been overlooked. This study developed a random forest (RF) model using data on 873 observations from 97 China construction companies in 2000-2017. Based on the variable importance analysis of RF, the top 10 variables were obtained and variables indicating both corporate governance and financial performance may be associated with an increased risk of corporate illegal activities. Then RF was compared with support vector machine (SVM) and the results indicate that both are suitable for predicting corporate misconduct in the construction industry. These findings expand the study of corporate misconduct in the construction industry and can be used to guide regulatory decision-making for conducting investigations into possible corporate misconduct.

[1] Ph.D. Student, Dept. of Civil and Environmental Engineering, Hong Kong Polytechnic Univ., 181 Chatham Road South, Hung Hom, Kowloon, Hong Kong SAR.
[2] Ph.D. Student, Dept. of Civil and Environmental Engineering, Hong Kong Polytechnic Univ., 181 Chatham Road South, Hung Hom, Kowloon, Hong Kong SAR.
[3] Associate Professor, Dept. of Civil and Environmental Engineering, Hong Kong Polytechnic Univ., 181 Chatham Road South, Hung Hom, Kowloon, Hong Kong SAR (corresponding author). E-mail: mark.hsu@ polyu.edu.hk
[4] Assistant Professor, Dept. of International Business, Tunghai Univ., No.1727, Sec.4, Taiwan Boulevard, Xitun District, Taichung, Taiwan.
[5] Distinguished Professor, Dept. of Civil Engineering, National Central Univ., No.300, Chung-da Rd., Taoyuan, Taiwan.

Keywords: corporate misconduct; random forest; support vector machine; variable importance

Introduction

Each year, dozens of deadly construction accidents occur worldwide. Many of these incidents are attributed to the large issue of corporate corruption. Corrupt practices have damaging consequences across multiple levels of the construction industry. For the local community, unemployment may rise, especially when the demand for related secondary business such as restaurants and gas stations decreases (Zahra et al. 2005). For society, the public's faith in senior managers and the ability of an executive board to monitor management is shaken (Zahra et al. 2005), with even confidence in the free market system eroded (Paruchuri and Misangyi 2015). This may cause a depressed moral climate in a society (Shadnam and Lawrence 2011). Apart from these repercussions, misconduct in the construction industry can lead to injuries and death. 11 workers were killed and 2 seriously injured after the collapse of an elevator at a Chinese construction site in April 2019 (Xinhua 2019).

Preventing such events is a top priority among practitioners and academics. A growing body of studies (Le et al. 2014; Liu et al. 2017; Owusu et al. 2019) have focused on identifying causal factors of corruption and generated numerous noteworthy factors. However, due to the limited budget and resources of a firm, coping with all those factors is very difficult. Even though a great deal of effort has been put into misconduct prevention practices and research, corporate scandals continue to arise. Therefore, it is essential to identify and rank the importance of possible factors. By focusing on the most important factors, investors, regulators, and other stakeholders could improve the effectiveness of misconduct detection and other critical evaluations.

Though recognizing those important risk factors could assist in mitigating corporate misbehaviors, timely and accurate detection of corporate illegal behaviors is also essential. However, accurately detecting corporate misconduct is a serious challenge. Some studies (Ngai et al. 2011; West and Bhattacharya 2016) claim that data mining approaches may be useful for detecting small anomalies because such approaches can extract and identify relevant information otherwise hidden in large volumes of data. Support vector machine (SVM) and other machine learning tools have been employed in analysis of construction cost, injury, contractor default, and other areas of the construction industry (Cao et al. 2014; Movahedian Attar et al. 2013; Tixier et al. 2016). The use of these tools, however, remains limited in the domain of construction corporate misconduct prediction. Wang et al. (2018) developed an SVM model to predict the occurrence of corporate misconduct in Taiwan based on several variables related to the board of directors. The study explored the role of statistically insignificant variables by comparing models with and without those variables, while also failing to provide a ranking of all variables, let alone the significant ones. In particular, when the number of factors is large, manual comparison would be time-consuming and inefficient. The present study draws upon a large quantity of data related to corporate governance and financial performance to rank feature importance and construct a data mining-based prediction model. By identifying the most influential factors, the prediction model is expected to provide regulators, investors and securities agencies with an effective and early misconduct detection tool.

Literature Review

Corporate Misconduct in the Construction Industry

Corporate misconduct is defined as the actions taken by companies to operate them illegally when they consider that the benefits outweigh the risks of doing so (Mishina et al. 2010). In the construction industry, various forms of misconduct have been identified, such as bid cutting (May et al. 2001), collusive tendering (Dorée 2004; Zarkada-Fraser and Skitmore 2000), and establishing front/shell companies (Chan and Owusu 2017). These behaviors may be attributed to underlying factors that are in play at different levels. From a macro perspective, flawed regulation systems may elevate the chances of opportunistic behaviors, and a negative industrial climate may encourage bad practices (Le et al. 2014). From a micro perspective, some scholars emphasize individual traits, like conducive attitude toward corruption (Brown and Loosemore 2015), egoism, and utilitarianism (Fan and Fox 2009). From the meso level, economic pressures (Alutu and Udhawuve 2009), board structure (Lee et al. 2018), organizational climate (Liu et al. 2017), commitment of code (Ameyaw et al. 2017), and other organizational factors may contribute to the occurrence of corporate misconduct. This study builds on the foundation of these organizational studies.

Although many factors have been identified as affecting the likelihood of corporate misconduct, less research considers ranking the importance of those factors and employing them to perform corporate misconduct prediction. Moreover, those studies investigating influencing factors relied on questionnaires, interviews, and other field survey tools to collect data. That is, the data sets are difficult to access by other researchers and the relationship between those underlying factors and corporate misconduct may not be verifiable. To address this gap, this study draws upon public information, especially from corporate annual reports, to serve as a proxy for organizational factors.

Random Forest

RF models have been used in various fields of science and engineering, including the construction industry. For instance, Tixier et al. (2016) developed a model to predict construction injury based on RF and Stochastic Gradient Tree Boosting with a set of features and safety outcomes extracted from textual injury reports. Liu et al. (2018) explored the impacts of outdoor ambient environment on scaffolding construction productivity via RF and a generalized additive model. Poh et al. (2018) presented an RF tool to explore safety leading indicators. Following this line of research, this study applies RF to corporate misconduct factor identification and prediction in the construction industry.

Random forest is an ensemble of small trees trained on a randomly selected sub-sample of a dataset through bootstrap aggregating or bagging (Breiman 1996). Each tree is trained through recursive partitioning of features to a certain level of depth, d. During this process, the randomly selected observations at each node are partitioned into subgroups to make a prediction (Breiman 2001). The exact partitioning position and the selection of features rely heavily on the distribution of observations (Strobl et al. 2009). The features, partitioning by which provides the most information regarding the observations, are chosen for this process. Several criteria are used for partitioning, but the most frequent ones are Gini Index (Breiman et al. 1984) for classification.

For each tree T_i $(i = 1,2, \dots, n_{tree})$, a new training data set S_i is generated by randomly resampling the original training data set $S = \{(x_i, y_i), i = 1,2, \dots, n\}, (X, Y) \in R^k \times R$. Although these sub-samples are different from each other, they must have similar distribution. Then tree T_i is created with the set S_i, by the above mentioned methodology and without pruning. In this process, some data will be used repeatedly while others might be "left out" and considered as out-of-bag (OOB) samples. This OOB data is used to evaluate the internal performance of each tree and to determine the variable importance (Breiman 2001). To increase the diversity of these trees further, m_{try} input variables are randomly selected from the k variables. Considering the m_{try} input variables and their linear combinations, a tree grows by searching the best split based on the generated training dataset and random variable set. In the same way, all the n_{tree} trees are constructed and trained. They are expected to be independent from each other because of the randomization of training data and input variables. Finally, all the constructed trees are collected into the RF model and vote for the outcomes.

For the sample x_t, $f(x_t) = majority\ vote\{T_i(x_t)\}_{i=1}^{n_{tree}}$ (1)

Corporate Misconduct Prediction

Though corporate misconduct prediction is not prevalent in the construction industry, some scholars have attempted similar prediction in the field of organizational management. Ravisankar et al. (2011) used a multilayer feed forward neural network, SVM, genetic programming, a group method of data handling, logistic regression (LR), and a probabilistic neural network to recognize fraud and non-fraud companies with 18 financial items. Pai et al. (2011) constructed an SVM-based fraud warning model to detect top management fraud based on 16 financial features about a firm's profitability, leverage, liquidity, and efficiency, as well as 2 variables about director shareholding. Lin et al. (2015) compared the performance of several data mining techniques (LR, DT, and Artificial Neural Networks) used as financial fraud detection tools with experts' judgments to analyze their differences. Most of the variables used were relevant to financial/accounting performance and several were relevant to corporate governance. Kim et al. (2016) established three multi-class prediction models using multinomial LR, SVM, and Bayesian networks. These models drew upon 49 variables, including off-balance sheet variables, nonfinancial measures, market variables and governance measures. Dong et al. (2018) adopted LR, SVM, DT, and neural networks and leveraged 3 categories of financial ratios and language-based features for financial misstatement detection.

Regarding input variables, most previous research employed financial/accounting variables. This may be related to the reasons for engaging in corporate misconduct. Unusual financial ratio values may represent a need to hide losses, to improve apparent stock market performance, and to satisfy investors, and lenders so as to mitigate managerial pressure (Ravisankar et al. 2011). Therefore, poor financial performance could be an incentive to commit corporate fraud. Fraud has been found to be conducted more often by top management (Zahra et al. 2005). As the chief decision makers, executives have the responsibility for setting the overall direction of an organization (Hambrick and Mason 1984). Once they decide how to behave, corresponding proper or improper actions within the firm follow. Thus, an array of studies attribute corporate fraudulent behaviors to the characteristics of top management (Schnatterly et al. 2018; Shi et al. 2016; Troy et al. 2011). In an effort to reduce such behaviors by executives, a board of directors is appointed by a firm's owners to serve as a monitoring device (Fama and Jensen 1983). A board of directors can play an important role in supervising and guarding against opportunistic behaviors by top management. The effectiveness of this function is associated with board size, board independence, and other board properties (Lee et al. 2018; Raheja 2005). Taken together, this may be why some studies (e.g., Kim et al. 2016; Pai et al. 2011) add several corporate governance related variables (e.g., CEO bonus and board shareholding) as input features. We followed the above studies and included variables about corporate governance and financial/accounting variables as our input features. Then, we ranked their importance, a step not typically considered in previous research, to identify the most influential factors of corporate misconduct in the construction industry.

As for classification techniques, previous studies have often used LR, SVM, and DT to develop their financial statement fraud detection models. Among them, LR is typically used as a benchmark (Ngai et al. 2011; Tserng et al. 2011). Though LR is easy to implement, it has difficulty in handling complex issues, especially fraud detection (West and Bhattacharya 2016). SVM is one of the most popular machine learning tools. It transforms the original data into a high dimensional space by nonlinear mapping and separates the data with a hyperplane. However, SVM is prone to overfitting (Pai et al. 2011). More importantly, SVM lacks variable importance ranking. With its ability to predict and provide variable importance, DT is an easy-to-use predictive model that generates mapping from observations to possible consequences (Ngai et al. 2011). It is constructed as a tree-like structure with attributes as branches and outcomes as leaves. When developing a predictive model, DT has no requirement for prior domain knowledge, making its implementation simple (Dutta et al. 2017). However, DT may be unstable and risks overfitting if a single tree is used (Bhattacharyya et al. 2011).

To overcome this drawback of DT, random forests (RF) was introduced by Breiman (2001). As an ensembled tool, RF is composed of a set of trees generated by a classification and regression tree (CART) (Breiman et al. 1984) and a combination of randomly chosen explanatory factors. This method inherits several advantages of DT (Sutton 2005). First, RF is able to handle complex nonlinear high-order interactions among features and does not require feature selection. It is also robust even with outliers and irrelevant inputs, as well as able to avoid overfitting (Rodriguez-Galiano et al. 2012). Next, there is no requirement for prior knowledge of underlying processes and no assumptions about the target function (Prinzie and Van den Poel 2008). RF has been shown to be among the most accurate general-purpose tools to date (Biau 2012). It additionally provides useful estimates of variable importance (Breiman 2001). With identifying variable importance and establishing an accurate prediction model as the primary aims of this study, RF is thus applied to the factor identification and prediction of corporate misconduct in the construction industry.

Method

Variable Importance

One of the most desirable characteristics of RF is its ability to generate variable importance. To compute the importance of a variable, RF first randomly permutes the value of a variable and keeps the others unchanged. Then a set of new trees is established. A set of accuracies corresponding to the modified OOB data is generated and compared with accuracies corresponding to the original OOB data with all of the variables. Their differences are calculated and averaged. The average value indicates the importance of that permuted variable. The larger the absolute value of the average of the differences is, the more important that variable is. The underlying rationale is that the data permutation of a variable would break its association with the output, and as a result, there would be a decrease in the accuracy if the permuted data were used as an input (Strobl et al. 2009). That is, if there is indeed a relationship between a variable and the output, replacing the original data with the permuted data would lead to a significant decrease in the accuracy, otherwise the replacement would make no difference to the accuracy. By doing so, RF reveals the variable importance and the association with the output. In particular, this association takes into consideration interactions with other variables (Strobl et al. 2009; Tsanas and Xifara 2012). The redundant variables are not given a priority even if they have a high correlation with the output. This function of RF facilitates research with high-dimensional data as is the case with the present study analyzing dozens of variables about financial performance and corporate governance.

Evaluation Metrics

Some studies (Bhattacharyya et al. 2011; Hajek and Henriques 2017) claim the cost of misidentifying lawful corporate behaviors as wrongful is much higher than that of neglecting to identify wrongful behaviors. This present study proposes that the cost of incorrectly classifying a lawful company as a violating one should not be overlooked as well. When a company is considered violating, subsequent investigation can be undertaken. If such actions are wasted on a lawful company, a fraudulent company would remain at large because of the limited resources of regulators. Moreover, investors would prefer to identify a trustworthy firm than a questionable one to achieve profits from their investments. Therefore, this study attempts to assess the performance of RF on both violating and lawful observations.

Whether the evaluated company is violating or lawful, the metrics used in this study are calculated mainly on the basis of the confusion matrix shown in Fig. 1.

Insert Figure 1 about here.

If the aim is to evaluate the performance of RF on violating observations, the violating companies are considered as positive while the lawful ones would be negative. Then TP is the number of violating observations classified correctly as violating. FN is the number of violating observations classified incorrectly as lawful. FP is the number of lawful companies falsely classified as violating while TN is the number of lawful companies accurately classified as lawful. On the other hand, if the aim is to evaluate the performance of RF on lawful companies, then the lawful companies are considered as positive while the violating one would be negative. TP and FN are the number of lawful observations correctly classified as lawful and wrongly classified as violating, respectively. FP and TN are the number of violating companies incorrectly classified as lawful and rightly classified as violating, respectively.

Based on the above confusion matrix, the metrics applied in this study include accuracy, precision, recall, and F1-score. These metrics can be formulated as follows:

$$Accuracy = \frac{TP+TN}{P+N} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{P} \tag{3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

Sample and Data

Our samples consist of all the publicly traded construction companies listed on the Shenzhen Stock Exchange and Shanghai Stock Exchange in China. All of these companies' information is derived from the China Stock Market and Accounting Research (CSMAR) database. This database collects financial and governance data mainly from the companies' annual, semi-annual, and quarterly reports. Some governance data is complemented by interim announcements by the board of directors, board of supervisors, and shareholder meetings. Regarding violation information, a list of violating companies was extracted from enforcement information published by the China Securities and Regulatory Commission (CSRC). By examining the violating cases carefully, this study identifies the year when violating behaviors are actually taken. If an illegal activity lasts for several years, we treat the company as a violator each year on the assumption that the activity could have been stopped at any time. If the date when a firm participated in fraud is not mentioned in the violating cases, it is assumed that the violation was detected immediately after the action took place. Though the CSMAR database collects enforcement information from 1994 to date, most records about construction companies begin after 2000. Thus, this study focuses on 97 construction companies over the period 2000-2017 to capture as much available data as possible. After data points with missing values were excluded, 873 final observations are yielded. Among them, 155 observations engaged in misconduct have been reported.

Measurement

As the output, corporate misconduct is operationalized by a dummy variable indicating whether an observation engaged in corporate misconduct or not. If yes, the observation is considered as violating and its label equals 1. Otherwise the observation is considered as lawful and its label is 0. This study employed 61 variables as the input, shown in Table 1. Among them, 24 were about corporate governance and the remaining were financial variables. These variables were selected because they encompass a wide cross-section of corporate governance information and financial ratios. Governance variables (X0-X23) show the structure, compensation, and other related information about the board and TMT. They have been reported to be related to illegal corporate behaviors (Chen et al. 2006; Dechow et al. 1996; Harris 2008; Jia et al. 2009; Kesner et al. 1986; Lee et al. 2018; Schnatterly et al. 2018; Sen 2007; Wowak et al. 2015; Zahra et al. 2005). Financial ratios included several financial aspects of the construction companies, i.e., structure ratios (X24-X28), liquidity ratio (X29-X36), growth capability (X37), operating capacity (X38-X46), per share indexes (X47-49), and profitability capacity (X50-X60). The financial variables were adopted mainly based on previous studies on fraudulent statement detection (Dutta et al. 2017; Hajek and Henriques 2017; Kim et al. 2016; Kirkos et al. 2007; Lin et al. 2015; Pai et al. 2011; Perols 2011; Ravisankar et al. 2011). Their calculation was based on the definition of CSMAR. Table 2 gives the descriptive statistics of the 61 variables.

Insert Table 1 about here.

Insert Table 2 about here.

Model Development

All the 893 observations were randomly and proportionally split into two parts. 80% were used as the training data (698 observations, 124 with corporate misconduct) while the other 20% were the testing data (175 observations, 31 with corporate misconduct). The training data was used to establish the learning model, and then the performance of the established model was evaluated adopting the testing data. All the variables were input without feature selection because of RF's ability to handle higher-order interactions among features.

Like other machine learning models, RF has several hyperparameters which need to be tuned (Breiman 2001; Ma and Cheng 2016). Previous studies (Poh et al. 2018) have mainly focused on the number of trees n_{tree} while other hyperparameters need to be meticulously tuned. In addition to the number of trees n_{tree}, the maximum depth which each tree will be split d, minimum number of samples on a node for branching S_n, minimum number of samples in a final leaf S_l, and features being considered for branching at each step $mtry$ are of equal importance. The sampling method could possibly affect the performance of RF. There is no effective method for simultaneous hyperparameter tuning of this model to the best of authors' knowledge. Therefore, grid search, a greedy search algorithm, was adopted for this study. In grid search, all possible initial values of hyperparameters are tested. Table 3 presents the list of hyperparameters and the search space of each one.

Insert Table 3 about here.

Each sample of the search space represented a possible set of hyperparameters. With each set, the dataset was randomly shuffled and the results of prediction were assessed with a 5-fold cross validation method. That is, 5 RF models were created and tested by splitting the dataset into 5 sections, and then, in 5 steps, keeping one part as the test set and the remaining as the training set. Their average was treated as the overall performance of that combination. Finally, the best candidate with the highest prediction accuracy was chosen as the hyperparameter set. These values are presented in Table 3. The processing time of this grid search by using scikit-learn, a library for machine learning algorithms with python (Pedregosa et al. 2011), took nearly 7.3 hours on a Core i7-8700T and 8.00 GB of RAM.

To assess the performance of RF further, a comparative analysis was conducted with SVM. SVM is commonly used in statement fraud detection, particularly in the construction industry. The same training and testing data with RF were scaled and inputted into SVM. In implementing SVM, two parameters were optimized, namely the penalty constant C and the radial basis function (RBF) kernel parameter g. They were also determined by grid search. That is, C and g were assigned a value from $\{2^{-10}, 2^{-9}, \ldots, 2^{9}, 2^{10}\}$ with 2^{1} as the exponential step. These combinations were tested by 5-fold cross-validation. In this study, the optimal C and g values were 64 and 0.0625, respectively.

Results and Discussion

1. Variable importance analysis

Variable importance as ranked by RF has the potential to facilitate the analysis of the role of input variables in corporate misconduct prediction. Fig. 2. depicts the following variables which are the most influential: ratio of net profits to total profits (X55), board of directors' total pay (X12), growth rate of total assets (X37), TMT total pay (X13), accounts payable turnover (X42), total pay for two boards and TMT (X11), current assets ratio (X24), net cash flow from operating activities per share (X49), ratio of total profits to EBIT (X56), and firm size (X2).

Among the top 10 features, 6 are associated with several categories of financial performance while the others are related to corporate governance. It is apparent that not only financial performance but corporate governance makes a significant difference in corporate misconduct prediction.

Insert Figure 2 about here.

The most important variable is ratio of net profits to total profits (X55), indicating the earnings capability of a firm. This capability is also represented by ratio of total profits to EBIT (X56), which is also among the top 10 variables. This shows that violating firms may try to inflate their profit or earning figures to create an impressive financial prospectus.

The second, fourth, and sixth important variables are board of directors' total pay (X12), TMT total pay (X13), and total pay for two boards and TMT (X11). All of them are associated with compensation. Regarding the designing and implementing total compensation package, compensation is a tool used by management for a variety of purposes to further the existence of the company. Directors with higher compensation are expected to contribute more to improving board effectiveness (Zhu et al. 2016). Effective board monitoring has been considered one of the most important mechanisms for preventing opportunistic managerial behaviors (Fama and Jensen 1983; Lee et al. 2018), such as corporate misconduct. Similarly, supervisors' compensation has been reported to be relevant to improving accounting information quality (Ran et al. 2015), which could be explained by supervisors with high compensation having a greater incentive to monitor directors and members of the TMT. TMT compensation, however, appears to operate differently than that of directors' and supervisors'.

High compensation may provide incentives to engage in fraudulent behaviors for executives to maximize their personal profits (Harris and Bromiley 2007). The tenth important variable is firm size. A larger firm is expected to have better internal governance and thus less likely to be involved in misconduct (Shan 2013). The ranking of these variables demonstrates the importance of corporate governance in preventing corporate misconduct.

The third important variable is growth rate of total assets (X37), reflecting a firm's growth capacity. Companies that are unable to achieve a certain performance level may be motivated to commit illegal activities to maintain their continuing growth (Harris 2008). The other important variables include a firm's operating capacity, ratio structure, and index per share, respectively. This indicates that any aspects of financial performance with an undesirable level may provide an incentive for corporate misconduct. Fortunately, those identified important variables serve to summarize comprehensive financial performance and thus improve the effectiveness of identifying questionable firms. The above results have important implications in the process of feature selection when establishing a corporate misconduct prediction model for construction companies.

2. Comparison between RF and SVM

According to the procedure described in model development, RF were trained, tested, and then compared with SVM to assess prediction performance. Table 4 shows the prediction results of RF and SVM. Their performance is very similar across all evaluation matrices. The accuracies of RF and SVM are both above 80%, indicating their overall performance is acceptable in predicting corporate misconduct. As we mentioned before, identifying both violating companies and lawful ones is meaningful. When predicting violating observations (label = 1), RF performs somewhat better than SVM in terms of precision (RF, 0.6667; SVM, 0.6250). The results show that RF identifies more actual violating observations than SVM among the observations labeled violating by the two algorithms. When predicting lawful companies (label = 0), the recall of RF (0.9931) is slightly higher than that of SVM (0.9792). This reflects that among all the actual lawful companies, more are identified by RF than SVM. In terms of overall performance, RF performs only a bit worse than SVM, with F1-scores and accuracy lower than those of SVM. This may be related to the high dimensionality of the dataset and correlated features, leading to the overfitting of SVM (Hajek and Henriques 2017; Pai et al. 2011). However, such a dataset and features won't affect the performance of RF. RF is robust even with high-order interactions among features, as mentioned in the literature review.

Insert Table 4 about here.

Moreover, both RF and SVM have higher precision, recall, and F-1 scores when the label is 0 than when the label is 1, showing that both perform better in identifying lawful observations than violating ones. This may be attributed to the fact that the number of violating observations is much smaller than that of lawful ones. Due to the somewhat limited sample size of violating companies, correctly predicting a violating company is more complex than predicting a lawful company using machine learning tools. As a result, it is difficult to precisely identify those violating companies. Nevertheless, accurately distinguishing lawful companies from those questionable ones is still meaningful. By giving those lawful companies an analog clearance certificate, the regulators could reduce the scale of investigation. Thus, the effectiveness of recognizing corporate misconduct may be subsequently improved. Simultaneously, investors could have greater confidence in their decision-making when selecting companies for investment.

Conclusion

Corporate misconduct can result in severe consequences, especially in the construction industry. Though previous studies have identified a great number of factors associated with corporate misconduct, ranking their importance and using them to predict corporate misconduct in the construction industry has been previously overlooked. To identify the most influential factors, this study developed an RF-based model employing a dataset about 873 observations from 97 China construction companies in 2000-2017. Among the 61 used variables, this study identified 10 variables, which represent several aspects of corporate governance and financial performance, with the greatest association with corporate misconduct. Then, based on the same dataset and inputs, the performance of RF was compared with that of SVM. The results show both are effective in predicting corporate misconduct of construction firms.

This study is expected to contribute to the field of corporate misconduct prediction. Using variable importance ranking of RF to explore the most influential factors, this study presents a method for locating key factors of corporate misconduct and for facilitating greater understanding of corporate misbehavior. In particular, the role of corporate governance deserves more attention in alleviating corporate misconduct. By employing RF and comparing it with SVM, this research demonstrates the feasibility of RF in predicting corporate misconduct in the Chinese construction industry. RF may provide a new option for researchers to more effectively identify questionable construction companies. This study also has practical implications. By exploring the most important factors, regulators and investors can be better equipped to more efficiently assess a firm's governance and financial condition and foresee the firm's possible behaviors. RF could be an effective tool for regulators and investors to identify both law-abiding and violating firms.

Though this research has included dozens of variables about corporate governance and financial performance, adding more features about projects, the firm itself, and its external environment may enhance the accuracy of corporate misconduct prediction in the construction industry. The variables used in this study were mainly extracted from a firm's annual reports, which also contain a textual description of a firm. Thus, combing for sentiment analysis with text mining tools could be helpful for identifying violating construction firms. The unsatisfactory performance of RF and SVM in predicting violating observations may be attributed to the imbalance in the data. The number of violating observations is far less than that of lawful observations. Supplementation with techniques addressing imbalance data issues would be beneficial. The RF model developed in this study uses data on Chinese construction firms only. Additional, similar research covering other industries and contexts is encouraged.

Data Availability Statement

All data and models used during the study are available from the corresponding author by request.

References

Alutu, O. E., and Udhawuve, M. L. (2009). "Unethical Practices in Nigerian Engineering Industries: Complications for Project Management." *Journal of Management in Engineering*, 25(1), 40–43.

Ameyaw, E. E., Pärn, E., Chan, A. P. C., Owusu-Manu, D.-G., Edwards, D. J., and Darko, A. (2017). “Corrupt Practices in the Construction Industry: Survey of Ghanaian Experience.” *Journal of Management in Engineering*, 33(6), 05017006.

Bhattacharyya, S., Jha, S., Tharakunnel, K., and Westland, J. C. (2011). “Data mining for credit card fraud: A comparative study.” *Decision Support Systems*, 50(3), 602–613.

Biau, G. (2012). “Analysis of a Random Forests Model.” *Journal of Machine Learning Research*, 13(4), 1063–1095.

Breiman, L. (1996). “Bagging predictors.” *Machine Learning*, 24(2), 123–140.

Breiman, L. (2001). “Random forests.” *Machine learning*, 45(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.

Brown, J., and Loosemore, M. (2015). “Behavioural factors influencing corrupt action in the Australian construction industry.” *Engineering, Construction and Architectural Management*, 22(4), 372–389.

Cao, M.-T., Cheng, M.-Y., and Wu, Y.-W. (2014). “Hybrid computational model for forecasting Taiwan construction cost index.” *Journal of Construction Engineering and Management*, 141(4), 04014089.

Chan, A. P. C., and Owusu, E. K. (2017). “Corruption Forms in the Construction Industry: Literature Review.” *Journal of Construction Engineering and Management*, 143(8), 04017057.

Chen, G., Firth, M., Gao, D. N., and Rui, O. M. (2006). “Ownership structure, corporate governance, and fraud: Evidence from China.” *Journal of Corporate Finance*, 12(3), 424–448.

Dechow, P. M., Sloan, R. G., and Sweeney, A. P. (1996). “Causes and Consequences of Earnings Manipulation: An Analysis of Firms Subject to Enforcement Actions by the SEC.” *Contemporary Accounting Research*, 13(1), 1–36.

Dong, W., Liao, S., and Zhang, Z. (2018). “Leveraging Financial Social Media Data for Corporate Fraud Detection.” *Journal of Management Information Systems*, 35(2), 461–487.

Dorée, A. G. (2004). “Collusion in the Dutch construction industry: An industrial organization perspective.” *Building Research & Information*, 32(2), 146–156.

Dutta, I., Dutta, S., and Raahemi, B. (2017). “Detecting financial restatements using data mining techniques.” *Expert Systems with Applications*, 90, 374–393.

Fama, E. F., and Jensen, M. C. (1983). “Separation of ownership and control.” *The Journal of Law & Economics*, 26(2), 301–325.

Fan, L. C., and Fox, P. W. (2009). “Exploring factors for ethical decision making: Views from construction professionals.” *Journal of Professional Issues in Engineering Education and Practice*, 135(2), 60–69.

Hajek, P., and Henriques, R. (2017). “Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods.” *Knowledge-Based Systems*, 128, 139–152.

Hambrick, D. C., and Mason, P. A. (1984). “Upper echelons: The organization as a reflection of its top managers.” *Academy of Management Review*, 9(2), 193–206.

Harris, J., and Bromiley, P. (2007). “Incentives to cheat: The influence of executive compensation and firm performance on financial misrepresentation.” *Organization Science*, 18(3), 350–367.

Harris, J. D. (2008). “Financial Misrepresentation: Antecedents and Performance Effects.” *Business & Society*, 47(3), 390–401.

Jia, C., Ding, S., Li, Y., and Wu, Z. (2009). “Fraud, enforcement action, and the role of corporate governance: Evidence from China.” *Journal of Business Ethics*, 90(4), 561–576.

Kesner, I. F., Victor, B., and Lamont, B. T. (1986). “Board Composition and the Commission of Illegal Acts: An Investigation of Fortune 500 Companies.” *Academy of Management Journal*, 29(4), 789–799.

Kim, Y. J., Baik, B., and Cho, S. (2016). “Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning.” *Expert Systems with Applications*, 62, 32–43.

Kirkos, E., Spathis, C., and Manolopoulos, Y. (2007). “Data Mining techniques for the detection of fraudulent financial statements.” *Expert Systems with Applications*, 32(4), 995–1003.

Le, Y., Shan, M., Chan, A. P., and Hu, Y. (2014). “Investigating the causal relationships between causes of and vulnerabilities to corruption in the Chinese public construction sector.” *Journal of Construction Engineering and Management*, 140(9), 05014007.

Lee, C. J., Wang, R., Lee, C. Y., Hung, C. C. W., and Hsu, S. C. (2018). “Board structure and directors’ role in preventing corporate misconduct in the construction industry.” *Journal of Management in Engineering*, 34(2), 04017067.

Lin, C.-C., Chiu, A.-A., Huang, S. Y., and Yen, D. C. (2015). “Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts’ judgments.” *Knowledge-Based Systems*, 89, 459–470.

Liu, J., Zhao, X., and Li, Y. (2017). “Exploring the Factors Inducing Contractors’ Unethical Behavior: Case of China.” *Journal of Professional Issues in Engineering Education and Practice*, 143(3), 04016023.

Liu, X., Song, Y., Yi, W., Wang, X., and Zhu, J. (2018). “Comparing the Random Forest with the Generalized Additive Model to Evaluate the Impacts of Outdoor Ambient Environmental Factors on Scaffolding Construction Productivity.” *Journal of Construction Engineering and Management*, 144(6), 04018037.

Ma, J., and Cheng, J. C. P. (2016). “Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests.” *Applied Energy*, 183, 193–201.

May, D., Wilson, O., and Skitmore, M. (2001). “Bid cutting: an empirical study of practice in South-East Queensland.” *Engineering, Construction and Architectural Management*, 8(4), 250–256.

Mishina, Y., Dykes, B. J., Block, E. S., and Pollock, T. G. (2010). “Why ‘good’ firms do bad things: The effects of high aspirations, high expectations, and prominence on the incidence of corporate illegality.” *Academy of Management Journal*, 53(4), 701–722.

Movahedian Attar, A., Khanzadi, M., Dabirian, S., and Kalhor, E. (2013). “Forecasting contractor’s deviation from the client objectives in prequalification model using support vector regression.” *International Journal of Project Management*, 31(6), 924–936.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011). “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature.” *Decision Support Systems*, 50(3), 559–569.

Owusu, E. K., Chan, A. P. C., and Shan, M. (2019). “Causal Factors of Corruption in Construction Project Management: An Overview.” *Science and Engineering Ethics*, 25(1), 1–31.

Pai, P. F., Hsu, M. F., and Wang, M. C. (2011). “A support vector machine-based model for detecting top management fraud.” *Knowledge-Based Systems*, 24(2), 314–321.

Paruchuri, S., and Misangyi, V. F. (2015). “Investor perceptions of financial misconduct: The heterogeneous contamination of bystander firms.” *Academy of Management Journal*, 58(1), 169–194.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research*, 12(Oct), 2825–2830.

Perols, J. (2011). “Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms.” *AUDITING: A Journal of Practice & Theory*, 30(2), 19–50.

Poh, C. Q. X., Ubeynarayana, C. U., and Goh, Y. M. (2018). “Safety leading indicators for construction sites: A machine learning approach.” *Automation in Construction*, 93, 375–386.

Prinzie, A., and Van den Poel, D. (2008). “Random Forests for multiclass classification: Random MultiNomial Logit.” *Expert Systems with Applications*, 34(3), 1721–1732.

Raheja, C. G. (2005). “Determinants of Board Size and Composition: A Theory of Corporate Boards.” *Journal of Financial and Quantitative Analysis*, 40(2), 283–306.

Ran, G., Fang, Q., Luo, S., and Chan, K. C. (2015). “Supervisory board characteristics and accounting information quality: Evidence from China.” *International Review of Economics & Finance*, 37(Supplement C), 18–32.

Ravisankar, P., Ravi, V., Raghava Rao, G., and Bose, I. (2011). “Detection of financial statement fraud and feature selection using data mining techniques.” *Decision Support Systems*, 50(2), 491–500.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. P. (2012). “An assessment of the effectiveness of a random forest classifier for land-cover classification.” *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104.

Schnatterly, K., Gangloff, K. A., and Tuschke, A. (2018). “CEO wrongdoing: A review of pressure, opportunity, and rationalization.” *Journal of Management*, 44(6), 2405–2432.

Sen, P. K. (2007). “Ownership Incentives and Management Fraud.” *Journal of Business Finance & Accounting*, 34(7–8), 1123–1140.

Shadnam, M., and Lawrence, T. B. (2011). "Understanding Widespread Misconduct in Organizations: An Institutional Theory of Moral Collapse." *Business Ethics Quarterly*, 21(03), 379–407.

Shan, Y. G. (2013). "Can Internal Governance Mechanisms Prevent Asset Appropriation? Examination of Type I Tunneling in China." *Corporate Governance: An International Review*, 21(3), 225–241.

Shi, W., Connelly, B. L., and Sanders, W. G. (2016). "Buying bad behavior: Tournament incentives and securities class action lawsuits." *Strategic Management Journal*, 37(7), 1354–1378.

Strobl, C., Malley, J., and Tutz, G. (2009). "An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests." *Psychological Methods*, 14(4), 323–348.

Sutton, C. D. (2005). "Classification and Regression Trees, Bagging, and Boosting." *Handbook of Statistics*, Data Mining and Data Visualization, C. R. Rao, E. J. Wegman, and J. L. Solka, eds., Elsevier, 303–329.

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

Troy, C., Smith, K. G., and Domino, M. A. (2011). "CEO demographics and accounting fraud: Who is more likely to rationalize illegal acts?" *Strategic Organization*, 9(4), 259–282.

Tsanas, A., and Xifara, A. (2012). "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools." *Energy and Buildings*, 49, 560–567.

Tserng, H. P., Lin, G.-F., Tsai, L. K., and Chen, P.-C. (2011). "An enforced support vector machine model for construction contractor default prediction." *Automation in Construction*, 20(8), 1242–1249.

Wang, R., Lee, C. J., Hsu, S. C., and Lee, C. Y. (2018). "Corporate misconduct prediction with support vector machine in the construction industry." *Journal of Management in Engineering*, 34(4), 04018021.

West, J., and Bhattacharya, M. (2016). "Intelligent financial fraud detection: A comprehensive review." *Computers & Security*, 57, 47–66.

Wowak, A. J., Mannor, M. J., and Wowak, K. D. (2015). "Throwing caution to the wind: The effect of CEO stock option pay on the incidence of product safety problems." *Strategic Management Journal*, 36(7), 1082–1092.

Xinhua. (2019). "11 killed, 2 injured in China construction site accident." Accessed on May 21, 2019. http://www.xinhuanet.com/english/2019-04/26/c_138009873.htm.

Zahra, S. A., Priem, R. L., and Rasheed, A. A. (2005). "The antecedents and consequences of top management fraud." *Journal of Management*, 31(6), 803–828.

Zarkada-Fraser, A., and Skitmore, M. (2000). "Decisions with moral content: collusion." *Construction Management and Economics*, 18(1), 101–111.

Zhu, J., Ye, K., Tucker, J. W., and Chan, K. (Johnny) C. (2016). "Board hierarchy, independent directors, and firm value: Evidence from China." *Journal of Corporate Finance*, 41, 262–279.

Fig. 1. Confusion matrix

Fig. 2. Importance ranking of variables

Table 1 Summary of input variables.

Variable	Description
X0: Capital structure change	Whether there is any change in the company's equity structure during the reporting period. 1 = unchanged, 2 = changed
X1: Relationship of top 10 shareholders	Three dummy variables representing whether top 10 shareholders are unrelated, related, or unconfirmed
X2: Firm size	Number of employees
X3: CEO duality	Whether the board chairman holds the managerial position CEO or president:1 = yes, 2 = no
X4: Board of directors' size	Number of directors
X5: Board independence	Number of independent directors
X6: Board of supervisors' size	Number of supervisors
X7: TMT size	Number of executives
X8: Board of directors' ownership	Number of shares held by board of directors
X9: Board of supervisors' ownership	Number of shares held by board of supervisors
X10: TMT ownership	Number of shares held by executives
X11: Total pay for two boards and TMT	Total annual emolument of directors, supervisors, and executives
X12: Board of directors' total pay	Total emolument of top 3 directors
X13: TMT total pay	Total annual emolument of top 3 executives
X14: Directors, supervisors, and executives with no salary	Number of directors, supervisors, and executives not receiving emolument
X15: Directors with no salary	Number of directors not receiving emolument
X16: Supervisors with no salary	Number of supervisors not receiving emolument
X17: Board committees	Total number of committees established
X18: The four board committees	Number of audit commission, strategic commission, nomination commission, and remuneration and evaluation commission established
X19: Other board committees	Number of other commissions established
X20: Working places consistency	Three dummy variables representing whether independent directors work in the same, different or unconfirmed place with the firm. When the number of independent directors is zero, the value is null
X21: Directors' meetings	Number of board of directors meetings

X22: Supervisors' meetings	Number of board of supervisors meetings
X23: Shareholders' meetings	Number of shareholder meetings
X24: Current assets ratio	Total current assets / total assets
X25: Ratio of working capital	(Current assets - current liabilities) / current assets
X26: Fixed assets ratio	Net fixed assets / total assets
X27: Ratio of shareholders' equity to fixed assets	Shareholders' equity/net fixed assets
X28: Current liabilities ratio	Total current liabilities / total liabilities
X29: Current ratio	Current assets / current liabilities
X30: Quick ratio	(Current assets – inventories) / current liabilities
X31: Times interest earned	(Net profits + income tax + financial expenses) / financial expenses
X32: Net cash flow from operating activities / current liabilities	Net cash flow from operating activities / total current liabilities
X33: Ratio of debt to assets	Total liabilities / total assets
X34: Ratio of long-term borrowings to total assets	Fixed assets / operating income
X35: Ratio of liabilities to tangible assets	(Total liabilities) / (total assets - net intangible assets - net goodwill)
X36: Ratio of equity to debt	Total owners' equity / total liabilities
X37: Growth rate of total assets	(Ending total assets - beginning total assets) / beginning total assets
X38: Ratio of accounts receivable to income	Accounts receivable / operating income
X39: Accounts receivable turnover	Operating income / ending accounts receivable
X40: Ratio of inventories to income	Inventories / operating income
X41: Inventories turnover	Operating costs / ending inventories
X42: Accounts payable turnover	Operating costs / ending accounts payable
X43: Current asset turnover	Operating income / ending balance of current assets
X44: Ratio of fixed assets to income	Fixed assets / operating income
X45: Fixed asset turnover	Operating income / ending balance of net fixed assets
X46: Total assets turnover	Operating income / ending balance of total assets
X47: Earnings per share	Net profits / ending paid-in capital
X48: Net assets per share	Ending owners' equity at period-end / ending paid-in capital
X49: Net cash flow from operating activities per share	Net cash flow from operating activities / ending paid-in capital
X50: Return on assets	Net profits / balance of total assets
X51: Net profits margin of current assets	Net profits / balance of current assets
X52: Net profits margin of fixed assets	Net profits / balance of fixed assets
X53: Return on equity	Net profits / balance of shareholders' equity
X54: Earnings before interest and tax (EBIT)	Net profits + income tax expense + financial expenses
X55: Ratio of net profits to total profits	Net profits / total profits

X56: Ratio of total profits to EBIT	Total profits / EBIT
X57: Ratio of EBIT to total assets	EBIT / total assets
X58: Gross operating margin	(Operating income - operating costs) / operating income
X59: Selling expense ratio	Selling expenses / operating income
X60: Operating margin before interest and taxes	(Net profits + income tax expense + financial expenses) / operating income

Table 2. Descriptive statistics (Mean ± St. Dev.) on financial variables

Variable	Mean ± St. Dev.	Variable	Mean ± St. Dev.
X0	1.6±0.49	X31	5.15±90.72
X1	2.38±0.61	X32	0.01±0.39
X2	14012.61±46830.75	X33	0.61±0.21
X3	1.82±0.38	X34	0.06±0.09
X4	9.03±2.02	X35	0.64±0.24
X5	3.16±0.97	X36	1.15±2.5
X6	3.86±1.23	X37	0.26±0.64
X7	7.41±3.3	X38	0.37±0.81
X8	45083015.2±129657700.54	X39	10±46.31
X9	758886.74±2416806.31	X40	0.56±1.55
X10	13762504.74±51422956.64	X41	11.63±53.33
X11	4382752.79±3968282.48	X42	4.29±5.28
X12	1295546.33±1068322.87	X43	0.95±0.54
X13	1361851.89±1096064.19	X44	0.43±1.13
X14	3.67±3.38	X45	24.83±277.42
X15	2.26±2.32	X46	0.61±0.33
X16	1.31±1.36	X47	0.31±0.5
X17	3.34±1.43	X48	3.9±2.63
X18	3.3±1.42	X49	0.21±1.36
X19	0.04±0.2	X50	0.02±0.18
X20	1.41±0.77	X51	0±0.5
X21	9.59±3.96	X52	-13.53±630.38
X22	5.26±2.33	X53	0.06±0.7
X23	2.99±1.63	X54	1250994070.05±5006172964.32
X24	0.67±0.21	X55	0.8±0.4
X25	0.15±0.24	X56	0.87±1.22
X26	0.14±0.14	X57	0.04±0.19
X27	87.55±1586.87	X58	0.17±0.14
X28	0.87±0.15	X59	0.02±0.03
X29	1.59±1.78	X60	0.06±0.82
X30	1.13±1.67		

Table 3. Results of hyperparameters tuning

Hyperparameter	Value	Search Space
n_{tree}	100	[50, 100, 150, 200, 250, 300,..,1000]
d	5	[3, 5, 7, …., 21] + [None]
S_n	2	[1, 3, 5, 7, 10]
S_l	1	[1, 3, 5, 7, 10],
$mtry$	All features	[Sqrt (features), Log_2(features), All features]
Sampling Method	Bootstrap	With/Without Bootstrap (sampling with replacement)

Table 4. Summary of prediction performance of RF and SVM

	RF		**SVM**	
Label	1	0	1	0
Precision	0.6667	0.8314	0.6250	0.8443
Recall	0.0645	0.9931	0.1613	0.9792
F1-Score	0.1176	0.9051	0.2564	0.9068
Accuracy	82.8571%		83.4286%	