https://dx.doi.org/10.1016/j.jclepro.2020.121395

Corporate environmental performance prediction in China: An empirical study of energy service companies

Saina Zheng[1], Chenhang He[2], and Shu-Chien Hsu[3,*], Joseph Sarkis[4], and Jieh-Haur Chen[5]

Abstract

Businesses are constrained by and dependent upon nature and institutional context. The global climate crisis has put pressure on and increased firm sensitivity to environmental issues. Predicting corporate environmental performance can help plan for environmental impact mitigation by adjusting organizational practices. Lack of environment-related information makes it difficult to make such predictions. A theoretical framework informed by the natural-resource-based view (NRBV) of the firm and institutional theory is used to identify variables for predicting corporate environmental performance. Five dimensions including institutional context, governance capability, information management capability, system capability, and technology-related capability, populated with 14 variables are used to empirically investigate the relationship of these variables with corporate environmental performance. Using 1100 data points on energy service companies (ESCOs) from 2011 to 2015 in mainland China, the Extreme Gradient Boosting (XGBoost) algorithm, a statistical nonlinear machine learning approach, is utilized to predict corporate environmental performance. The results demonstrate that the XGBoost model can be effective for ESCO environmental performance prediction, with satisfactory prediction accuracy. This study also adopted the SHapley Additive

[1] Ph.D. candidate, Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University.
[2] Mphill candidate, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University.
[3] *Corresponding author: Assistant Professor, Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University.
E-mail: mark.hsu@polyu.edu.hk; Tel: +852-27666057
[4] Professor, Foisie Business School, Worcester Polytechnic Institute.
[5] Distinguished Professor, Department of Civil Engineering, National Central University.

exPlanations (SHAP) values for model interpretation, indicating that total assets, amount of proactive environmental costs, proportion of technicians and number of patents contribute most to corporate environmental performance. Several policies and environmental strategies for improving corporate environmental performance in the ESCO industry are derived from this analysis.

Keywords: Corporate environmental performance, Extreme Gradient Boosting (XGBoost), Energy service company (ESCO).

1 Introduction

Growing climate and other environmental crises, such as resource depletion, have led to an increased focus on shifting to a low-carbon world (Millar et al., 2018). Policies have focused on attempts to cut or mitigate greenhouse gas (GHG) emissions, as policy measures are the most direct way to reduce the risk of future climate change impacts (IPCC, 2015). Countries are joining global environmental collaborative efforts including the Kyoto Protocol, the Copenhagen Accord, and the Global Pact for the Environment. Several market-based environmental instruments including green credits, green insurance, and pollution tax policies have been adopted (Crowley, 2013; Garnaut, 2008; Neuhoff, 2011; Newell and Paterson, 2010; Nyberg et al., 2013; Stern, 2008). According to the Country Policy and Institutional Assessment (CPIA), ratings on policies and institutions for environmental sustainability of the world have continued to rise since 2005, as shown in Figure 1(a). Figure 1(b) indicates that countries with higher income tend to pay more attention to policies concerning environmental sustainability. With its rapid economic development, China is expected to launch more environmental policies.

(a) (b)

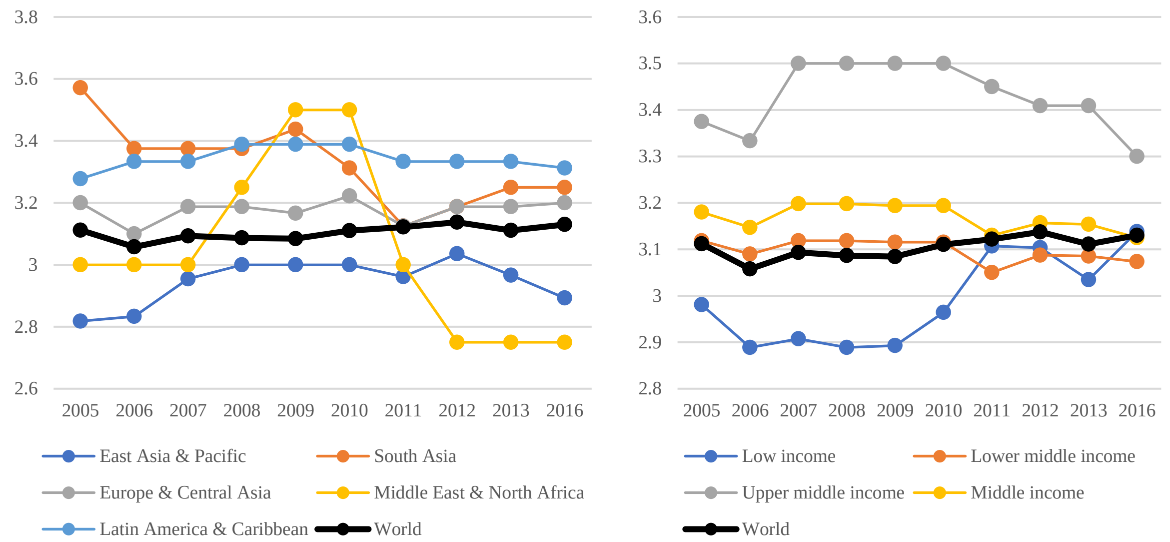

Figure 1: CPIA ratings on policy and institutions for environmental sustainability (1=low to 6=high). (a) CPIA ratings on policies and institutions for environmental sustainability by region; (b) CPIA ratings on policies and institutions for environmental sustainability according to country income level. (Data source: World Bank)

China became the world's largest carbon emitter in 2006 and has increased attention to environmental issues arising from its population growth and economic development (Liu et al., 2016; Zhang et al., 2008). Past environmental policy in China focused on mandatory regulations. In recent years this role has shifted to market-oriented and voluntary approaches. Fiscal incentive policies, tax subsidies, a pollution levy system, and technology innovation support are being provided seeking to achieve economic and environmental protection 'win-wins' (Zhang et al., 2007). Energy performance contracting (EPC) is one of these instruments. EPC is a market-oriented approach in which the energy service companies (ESCOs) invest in implementing energy services for customers to improve energy efficiency, including energy savings guarantees, associated design, and installation services. ESCOs get paid annually from energy savings during the contract period (Deng et al., 2017; Zheng et al., 2018). ESCOs are corporations which focus on improving energy efficiency and relieving climate change through EPC (Liu et al., 2018; Xu et al., 2015; Xu and Chan, 2013).

Economic growth has been identified as the main driver for sharp CO_2 emissions increases. Anthropogenic causes of climate change are intimately related to economic behaviour, and the industry is increasingly being called upon to respond (Yeeles, 2018). The sheer scale of the Chinese economy means that worldwide CO_2 emissions are strongly determined there (Wiedenhofer et al., 2017). The enterprises are the primary damager of environmental pollution and the major consumer of energy in China (Li et al., 2017). Nearly two-thirds of China's groundwater was of poor quality, over 15% of China's soil and farmland has been polluted, causing serious threat to food security and human health (Li et al., 2017; Qiu, 2011). To reduce these impacts, regulating corporate environmental performance in China is in urgent need.

The institutional theory stipulates that firms will respond to institutional pressures (mainly regulatory policy pressures) to thrive and gain legitimacy (Meyer and Rowan, 1977; Scott, 2013). Damaging corporate environmental impact has become increasingly part of the public and social mindset. This concern over climate change and the policy regulations to mitigate GHG emissions have exerted greater pressures on corporations to improve environmental well-being rather than hastening its degradation. These forces have also incentivized corporations to pursue good environmental performance (Hart, 2010).

The natural-resource-based view (NRBV) of the firm stipulates that a business will be constrained by natural resources and firm competitiveness is related to natural resources (Hart, 1995). Many corporations now claim that developing a corporate culture that promotes social and environmental sustainability can improve employee recruitment attraction, motivation, and retention (Renwick et al., 2013). Corporations are also realizing the significance of climate change and develop strategies for this environmental issue (Wright and Nyberg, 2015).

Business and industry play a dual role in climate politics. Firstly, corporations are the principal agents producing CO_2 emissions; secondly, corporations can improve the environment and reduce emissions through technological innovation. Better environmental performance reduces the volatility of the firm's cash flows, decreases potential bankruptcy costs, and increases debt capacity; all characteristics that can add resources for an organization to build competitive advantage. Many theoretical and empirical studies also indicate that better environmental performance boosts and is endogenously influenced by better financial performance (Dixon-Fowler et al., 2013; King and Lenox, 2001; Stanwick and Stanwick, 1998).

Disclosure of corporate environmental performance has been advocated to achieve better environmental performance. Scholars have argued that information about corporate environmental performance disclosed to the public can play an important role in determining business strategies, consumers' purchasing behaviour, and investors' financial investment decisions (Meng et al., 2014; Rockness, 1985; Spicer, 1978). Environmental disclosure may decrease the agency costs of debt and reduce estimation or information risk (Bansal and Clelland, 2004; Gao and Connors, 2011). However, a vast majority of companies do not produce corporate environmental reports or include environmental information in their annual reports. This result may be due to environmental information disclosure resistance, a desire to avoid additional costs, fear of threats to local employment, and concerns about reduced profits (Wang et al., 2004).

Researchers have been investigating corporate environmental performance for decades. They have mainly focused on evaluating environmental performance and environmental management strategy using the environmental-related information (Bhatnagar, 1999; Delmas and Blass, 2010; Ilinitch et al., 1998; Klassen and McLaughlin, 1996; Lober, 1996; Tyteca et al., 2002; Zhang et al., 2008). In China, various corporations have made efforts toward environmental protection and generating data related to environmental performance, especially given governmental pressures for this type of information. This data provides information which can be useful for practical governmental and organizational policy decisions, but also for research purposes. However, this type of data is currently rare in China, resulting in some hurdles in predicting corporate environmental performance using environment-related information. Only a few studies emphasize predicting environmental performance, due to the scant data available and lack of a detailed list of pollutants emitted by corporations (Delmas and Blass, 2010). It is essential find ways to utilize corporate related information, which can be accessed easily, for predicting corporate environmental performance. An application using machine learning method-XGBoost can help in completing various predictive analyses for multiple settings and purposes, especially when there exists the sparse and noise in the dataset.

In this study, a machine learning model for predicting corporate environmental performance is constructed with two main functionalities: assessing the corporate environmental performance of an unknown ESCO and calculating the future performance of ESCOs. Predicting corporate environmental performance can be used to mitigate environmental impacts through guiding organizational practices, and to improve a firm's reputation. First, we combine the elements of Institutional Theory and NRBV to bring a fresh perspective to environmental performance prediction research. Second, we explore the factors in five domains: Institutional context, Governance capability, Information management capabilities, Systems capability, and Technology-related investment, using 1100 data points on corporate environmental performance from different industries in mainland from 2011 to 2015 in China. Lastly, we utilize a machine learning tool, XGBoost regression to predict future performance and adopt Shapely to generate interpretations from the model. Conclusions and future research finalize the paper.

2 Theory foundation and hypotheses development

Corporate environmental performance can be defined as the results of an organisation's management of its environmental aspects or more precisely 'is the totality of a firm's behaviour toward the natural environment (i.e. it's level of total resource consumption and emissions)' (Tyteca et al., 2002). Corporations compete over limited natural resources, tend to take strategies to use the resources more efficiently, relieve their impact on the natural environment, and focus more effort on pollution control. Corporate environmental performance evaluation has been proposed for self-assessment, benchmarking, and reporting (Delmas and Blass, 2010; Gao and Connors, 2011; Ilinitch et al., 1998; Veleva and Ellenbecker, 2001).

Several theories have been used to investigate and explain corporate environmental performance. These theories include the natural-resource-based view (NRBV) (Hart, 1995), institutional theory (Colwell and Joshi, 2013; Jennings and Zandbergen, 1995), stakeholder theory (Freeman, 1984), agency theory (Berrone and Gomez-Mejia, 2009; Friedman, 2007), and transaction cost theory; to name a few organizational theories. Two of these are especially popular and salient. One is a general external to the organization theory, institutional theory, the other is an internal theory used to build competitive advantage, the NRBV. Together these two theories provide a more complete picture of how organizations manage their environmental performance.

2.1 Combining Institutional Theory and the Natural-Resource-Based View

The institutional theory posits that organizations enhance or seek to protect their legitimacy (Scott, 2013) by conforming to the expectations of institutional norms and stakeholder requirements (Aldrich and Fiol, 1994; DiMaggio and Powell, 2000). Concern over legitimacy forces firms to adopt managerial practices that are expected to conform to social values and expectations (Berrone and Gomez-Mejia, 2009). With the increasing importance of environmental issues, institutional theory stipulates that companies under heavier institutional pressure will gain legitimacy by exhibiting good environmental performance (Bansal and Clelland, 2004; Bansal Pratima, 2005). Researchers have applied institutional theory in the investigation of corporate environmental performance (Berrone and Gomez-Mejia, 2009; Campbell, 2007; Gallego-Alvarez et al., 2017; Tashman and Rivera, 2016). The institutional context has a significant influence on environmental performance and the adoption of environmental strategies (Chang et al., 2015; Christmann, 2004; Russo and Fouts, 1997; Sharfman et al., 2004; Wang et al., 2018). Under institutional pressures, firms have tended to adopt appropriate strategies and firms with an environmental legacy has incurred less risk (Bansal and Clelland, 2004).

NRBV (Hart, (1995) and holds that the business is constrained by and dependent upon natural ecosystems. Organizational competitiveness relies on the capabilities which facilitate environmentally sustainable economic activity. Many researchers have examined the relationship between corporate environmental performance and financial performance (McWilliams and Siegel, 2001; Stanwick and Stanwick, 1998). Al-Tuwaijri et al (2004) provided an analysis of the interrelations between environmental performance and economic performance, finding that good environmental performance is significantly associated with good economic performance. Trumpp and Guenther (2017) build on the theory of a non-linear, specifically a U-shaped, relationship between corporate environmental performance and corporate financial performance.

Empirical work by Buysse and Verbeke (2003) identified five essential resource domains through which environmental proactiveness can be determined: strategic environmental planning, formal routine-based environmental management, organizational competencies in environmental management, employees’ green skills, and conventional technology-based green competencies. These determinants have also been categorized into four resource domains: governance capability, information management capabilities, systems capability, and technology-related investment (Backman et al., 2017). These elements will prove helpful in our investigation using our machine learning models.

Scholars across a wide range of research areas and disciplines have focused on examining the relationship between corporate environmental performance and other organizational constructs or variables (Bansal and Gao, 2006; Trumpp et al., 2015). These variables include external organizational factors, such as regulation (Camisón, 2010) or stakeholder pressure (Ilinitch et al., 1998), as well as internal organizational factors, such as different characteristics of the board (Post et al., 2015) or innovation (Hall and Wagner, 2012). These findings corroborate the NRBV and institutional theory, dividing corporate environmental performance into five categories: institutional context, governance capability, information management capabilities, systems capability, and technology-related capability. We conceptualize how the possible combinatorial configurations from institutional theory and the NRBV relate to corporate environmental performance and try to understand how corporations may achieve improved environmental performance.

Figure 2 summarizes the integration of institutional theory and NRBV for corporate environmental performance. The figure depicts how corporations will choose to engage in environmentally friendly behaviour due to limited natural resources in a given institutional context. Institutional theory is adopted to explain how organizations react to institutional pressures, while NRBV encompasses building corporate capabilities in such a way to gain a competitive advantage in the market given natural resources consideration (Delmas and Toffel, 2004; Hart, 1995). Institutional theory categorizes the institutional pressure into three types, namely, cognitive, regulative and normative pressure (Gao et al., 2019). The cognitive pressure is related to the economic and ethical aspects which mainly refers to the environmental benefit and ethical obligation (Gao et al., 2019). The regulative pressure comes from the regulations, laws, rules and other formal instruments. The normative pressure is those pressure exerted by the nongovernmental stakeholders, such as suppliers, consumers, competitors (DiMaggio and Powell, 2000; Gao et al., 2019; Lee et al., 2018). Firms depend directly on natural capital and ecosystem services (Pogutz & Winn, 2009; Starik & Rands, 1995). Without air, water, a favourable climate, and a variety of natural resources, no organization can survive (Gladwin et al., 1995). Key resources and capabilities also affect organizational ability to adopt competitive environmental strategies (Hart, 1995). With the increasing consequences of climate change and growing severity of resource scarcity, firms are facing loss of access to natural resources and must adapt according to their dependence on nature. Both theses institutional pressures and natural resources pressure drive the organization to minimize emissions, effluents, waste, life-cycle environmental costs of products, and environmental burden of firm growth and development. That is, organizations adopt environmental strategies to achieve higher corporate environmental performance. Corporate capabilities are the key factors which will affect the adoption of environmental strategies. As stated by NRBV, these capabilities consist of governance capability, information management, systems, and technology-related capabilities. Corporates will adopt environmental strategies to pursue competitive advantages in the market since researchers found that firms with better environmental performance have superior financial performance (McWilliams et al., 2006). Environmental strategies can also lead to reduced costs and improved environmental performance simultaneously (Lyon and Maxwell, 1999).

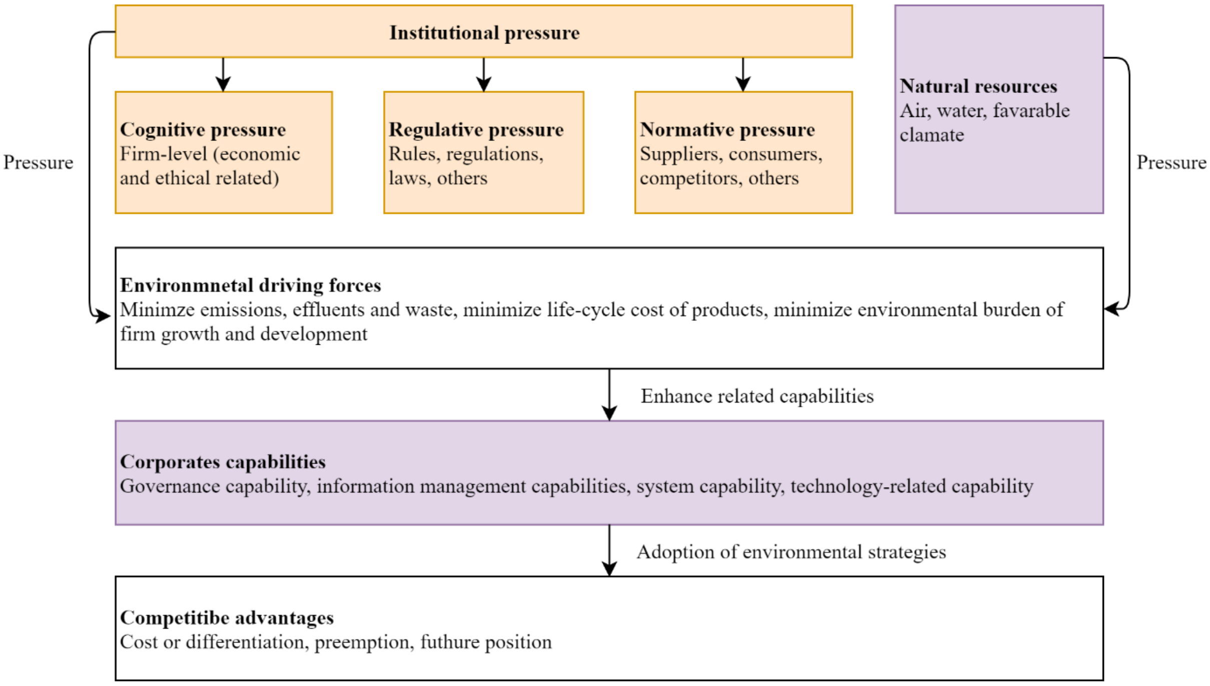

Figure 2: Framework of corporate environmental relationships and performance based on Institutional Theory and the Natural-Resource-Based View of the firm.

2.2 Institutional context and corporate environmental performance

Institutional forces play a significant role in corporate environmental strategy adoption (Chang et al., 2015); impacting corporate environmental performance. High-income regions of the world have generally exerted the strongest regional or national institutional pressure for improved environmental performance. Institutional pressure has been found to be lower in middle-income and lessens in lowest-income regions (Luxmore et al., 2018)). The reason for this is an overwhelming need for economic development in some regions, where institutional measures from an environmental perspective may be lessened.

In China, developed regions and their governmental agencies, invest more in improving energy efficiency and require organizations to emphasize environmental performance (Zheng et al., 2018). In this study, the gross domestic product (GDP) and the population are proxies for institutional (government) pressure since this metric reflects the development of a region. Natural resource availability may also exert pressure on corporate environmental performance from NRBV. Consumption of coal is used to reflect the level of dependency on natural resources; renewable energies may also be dependent on natural resources but are not as constrained due to the continuous sources (e.g. sunlight and wind power). Using these perspectives, the following hypothesis is proposed for testing.

Hypothesis 1 (H1): Corporates facing greater institutional and natural resource pressures will exhibit greater improvements in corporate environmental performance.

2.3 Organizational characteristics and corporate environmental performance

There are multiple levels of pressures and contexts. The first hypothesis focused on broader social and natural resource considerations and relationships to environmental performance. Organizational (corporate) contexts and characteristics will also relate to corporate environmental performance. Corporate specific internal resources and capabilities are particularly useful in generating unique, preventive and voluntary environmental actions to reduce firms' environmental impacts (Hart, 1995). According to NRBV, corporate characteristics can be divided into four domains: governance, information management, system, and technology-related capabilities.

(1) Governance capability indicators

Governance capability refers to a strategic planning process reconfiguration ability and integration of environmental issues into corporate policies and routines (Backman et al., 2017; Walls et al., 2012). A measure of governance capability and policy focus includes environmental costs which are the investments made in addressing pollution issues and adopting environment strategies (Salo, 2008);. Another, proxy measure includes the number of formal legal warnings a firm has received since its founding. This warnings measure indicates how well the governance structure supports good or poor behaviour and can be closely linked environmental policy; e.g. going beyond compliance (Li et al., 2017).

Firms with a high level of environmental commitment and stronger governance policies are more likely to regard environmental protection as their corporate social responsibility and be eager to protect the environment, thus achieving higher corporate environmental performance (Al-Tuwaijri et al., 2004; Muller and Kolk, 2010; Wang et al., 2018). Furthermore, organizations with historically poor environmental records are often subjected to more scrutiny by their local communities and regulators. Thus, organizations with poor environmental records may try to build greater corporate environmental governance capabilities to achieve higher environmental performance to gain more resources. Together, the following is expected.

Hypothesis 2 (H2): Greater organizational environmental governance capability, measured by the combination of environmental cost and formal legal warning, relates to higher corporate environmental performance.

(2) Information management capability

Information management capabilities mainly focus on formal management systems and procedures of investment. Researchers found that effective information management capabilities and corporate social responsibility are synergistically related, and can facilitate transition to corporate sustainability (Gangi et al., 2019). Countries paying attention to climate change mitigation tend to set develop stronger information management capability for organizations (Backman et al., 2017). This information refers to environmental-related information, such as the climate change impact mitigation and carbon footprint, denoting the attitude towards sustainability. The work environment is considered to evaluate the organizational culture regarding how a corporation views the importance of environment (Bhatnagar, 1999). High environmental awareness can help firms to implement environmental management practices smoothly and then help them improve environmental performance. Based on this analysis, the following is hypothesized.

Hypothesis 3 (H3): Corporates with stronger information management capability tend to achieve higher corporate environmental performance than corporates with weaker information management capability.

(3) Systems capability

Systems capability covers investments in employee skills and organizational competencies, such as research and development funding, finance and accounting, and storage and human resources in environmental management (Backman et al., 2017; Buysse and Verbeke, 2003). Previous research investigated the relationship between organizational characteristic variables and environmental performance/environmental benefits, such as the top management's leadership skills, human resources and organizational size (Etzion, 2007; Lee et al., 2018). Kitada and Ölçer (2015) put forward that employee element is essential when considering corporate social responsibility. It was found that there appears to be a positive relationship between a firm's environmental performance and its financial performance (Dixon-Fowler et al., 2013; Rockness, 1985; Spicer, 1978). Accordingly, we postulate the following.

Hypothesis 4 (H4): Corporations with stronger system capability tend to achieve higher corporate environmental performance than corporates with weaker system capability.

(4) Technology-related capability

Technology-related capability covers the conventional green competencies related to green product and manufacturing technologies. Technologies will affect corporate competitiveness since the environmental problems arise increasing awareness (Shrivastava, 1995). Technology in energy efficiency proved options and solutions for organizations to pursue better environmental performance by implementing the energy efficiency retrofit projects (Kitada and Ölçer, 2015). Benitez-Amado and Walczuch (2012) believed that technology-related capabilities are a key enabler for organizations to achieve better environmental performance. Environmental innovations contribute to corporate environmental performance since they can improve energy efficiency and reduce pollution. Therefore, our last hypothesis is as following (Kagan et al., 2003).

Hypothesis 5 (H5): Corporates with stronger technology-related capability tend to achieve higher corporate environmental performance than corporates with weaker technology-related capability.

3 Research method and data processing

3.1 Sample and data

The combination of institutional theory and NRBV identifies five domains for selecting the index to predict corporate environmental performance. Considering the data availability and referring to the previous research, 14 factors as shown in Table 1 were chosen to test the hypotheses. The research data we employed was provided by the ESCO Committee of China Energy Conservation Association (EMCA). The collected data covers 3225 ESCOs in 30 provinces in mainland China from 2011 to 2015 (Zheng et al., 2018). However, some ESCOs were excluded for one or more of the following reasons: (i) data for environmental performance is missing; (ii) data for more than 3 variables are missing. Thus, 1134 ESCO projects have sufficient information for further analysis. The value of the corporate environmental performance for most projects are between 0 and 1, however, the corporate environmental performance of 34 projects (3% of total projects) is 0, meaning there is no environmental income for these companies, which is not suitable for our research. Then, 1100 ESCO projects are finally analysed to predict the corporate environmental performance, which is mainly located in the Beijing, Shandong, and Guangdong provinces (see Fig. 3). Table 2 shows an example of the detailed information for each project, including investment, number of formal legal warnings since foundation, proportion of in-plant environmental, proportion of technicians, assets, equity, environmental projects payback period, asset age, revenue, tax bracket, and number of patents. All the variables in Table 1 can get or calculated based on

Table 2. The amount of proactive environmental costs is the investment for improving energy efficiency and reducing the impact of environment. The proportion of technician can be get using the number of technicians divided by number of employees. Information related to GDP, population, and consumption of coal was gained through the National Bureau of Statistics of China.

Table 1: Corporate environmental performance indicator system.

Destination layer	Standard layer	Index layer	Data source	References
Corporate environmental performance prediction indicator system	Institutional context	GDP (GDP)	National Bureau of Statistics of China	(Chan and Makino, 2007; Zheng et al., 2018)
		Population (PO)	National Bureau of Statistics of China	(Cui and Jiang, 2012)
		Consumption of coal (CC)	National Bureau of Statistics of China	(Zheng et al., 2018)
	Governance capability	Amount of proactive environmental costs (PEC)	EMCA	(Fu et al., 2017; Salo, 2008)
		Number of formal legal warnings since firm founding (FLW)	EMCA	(Li et al., 2017; Yoon et al., 2006)
	Information management capability	Proportion of In-plant environment (PIE)	EMCA	(Bhatnagar, 1999)
	Systems capability	Proportion of technicians (PT)	EMCA, $\frac{number\ of\ techinicians}{number\ of\ employees}$	(Etzion, 2007; Lee et al., 2018)
		Total assets (TA)	EMCA	(Backman et al., 2017; Buysse and

				Verbeke, 2003)
		Equity (EQ)	EMCA	(Backman et al., 2017; Buysse and Verbeke, 2003)
		Environmental projects payback period (PP)	EMCA	(Dibrell et al., 2011)
		Asset age (AA)	EMCA	(Li et al., 2017)
		Revenue (RE)	EMCA	(Orlitzky et al., 2003; Russo and Fouts, 1997)
		Tax bracket (TB)	EMCA	(Hoi et al., 2013)
	Technology-related capability	Number of patents (PA)	EMCA	(Benitez-Amado and Walczuch, 2012)

Table 2: Example of detailed information about ESCO

Liaoning Nengfaweiye Energy Technology Co., Ltd.	Region	Number of Employees	Number of Technicians	Number of Patents	Investment (million yuan)	Assets
	Liaoning	**450**	**68**	**13**	**27.53**	**240.23**
	Equity	Payback Period	Asset Age	Ratepaying Credit Grade	Number of Penalties Received	Environmental Performance (Environmental income per unit of an asset)
	20.205	**0.8**	**6**	**A**	**0**	**0.354502**

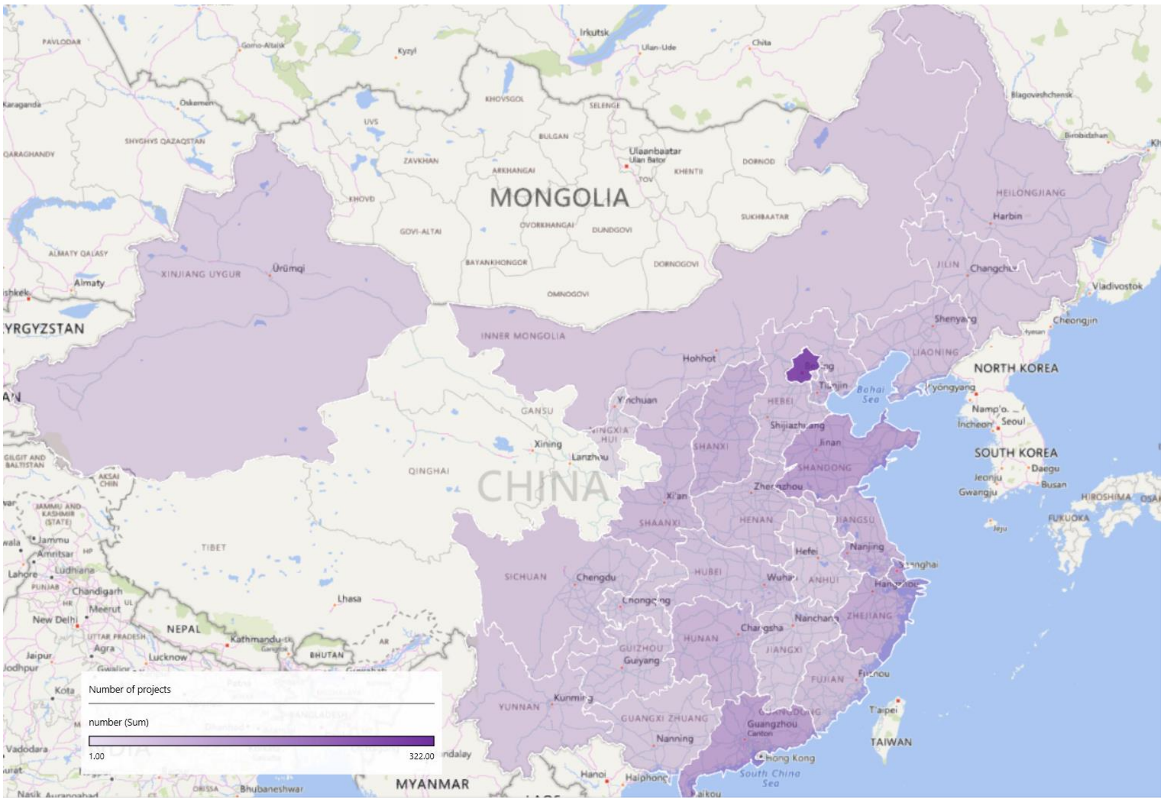

Figure 3: Distribution of Sampled ESCOs

These samples cover all kinds of firms, including state-owned enterprises, corporations, general partnership firms, private enterprises, foreign-owned enterprises, and others, with their assets varying from 0 to more than 1 trillion yuan (shown in Table 3).

Table 3: Sampled firms by business type and asset size

Business type	Number
State-owned enterprise	157
Corporation	128
General Partnership	54
Private enterprise	736
Foreign-owned enterprise	14
Other	11

Assets (in yuan)	Number
0-500	20
500-1000	202
1000-5000	531
5000-10000	171
10000-100000	171
≥100000	5

Fig. 4 displays the scatter plots for each of the (normalized) input variables and output variables. These scatter plots show that none of the functional relationships between the input variables and the output variables are trivial.

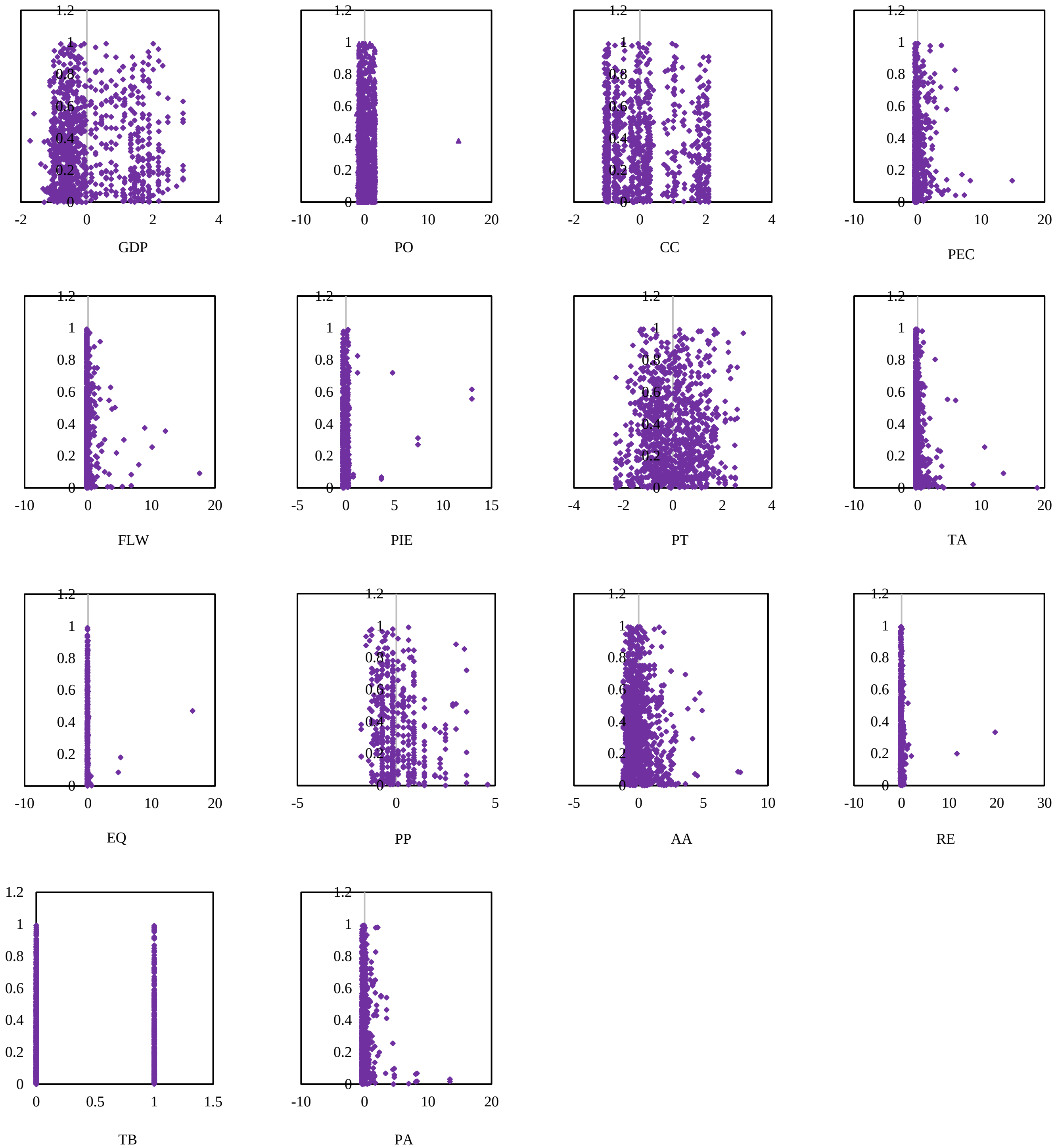

Figure 4: Scatter plots of the relationships between each input variable and output

This suggests that we can reasonably accept that classical learners such as linear regression may fail to find an accurate mapping of the input variables to the output variables. Therefore, these plots intuitively justify the need to experiment with more complicated learners such as machine learning methods. However, the machine learning methods are mainly used for prediction and classification, without the ability to interpret the relationship between variables. Recently, SHAP (SHapley Additive exPlanations) was developed to interpret the variables' impact on the model's prediction (Lundberg and Lee, 2017). A SHAP value for a feature of a specific prediction represents how much the model prediction changes when we observe that feature. This SHAP figure not only indicates which features are most important but also their range of effects over the dataset, revealing the relationship between variables and model output.

3.2 Machine learning method-XGBoost

In recent years, machine learning has been generating a lot of curiosity for its superior performance compared to other more traditional statistical techniques. Numerous machine learning models like Linear/Logistic regression, Support Vector Machines, Neural Networks, and Tree-based models are being tried and applied in analysis and prediction (Gumus and Kiran, 2017). Tso and Yau (2007) predicted electricity energy consumption adopting the decision tree and neural network models. Lee (2007) applied support vector machines to suggest a new model for corporate credit ratings with better explanatory power and stability. Among these methods, Extreme Gradient Boosting, also known as XGBoost (Chen and Guestrin, 2016), is a model that has a high success rate in the majority of machine learning competitions and has proven to be efficient for predictive modeling.

XGBoost has algorithms that can deeply explore data-label correlations by adaptively fitting large-scale data via tree boosting. Compared to conventional regression approaches such as logistic regression and SVM regression, XGBoost's tree-ensemble approaches can easily handle data with missing values (Torlay et al., 2017). Third, XGBoost penalizes the complexity of an individual tree as a regularization term, which has better generalization ability compared to other MART (multiple additive regress trees) methods. Ajit and Punnoose R (2016) applied XGBoost to predict employee turnover within an organization, addressing the prevalence of noise in data to reduce overfitting and improve accuracy. XGBoost is suitable for our case since there exist sparse data and noisy data in the realm of corporate environmental performance. Furthermore, the tree-ensemble algorithm provides strong interpretability of the model. By constructing the model, we can visualize the tree's structure and explore implicitly how the model makes decisions and which attributes are dominant.

XGBoost is a typical tree-ensemble model related to CART (Classification And Regression Trees), which grows the tree in a top-down manner. Each tree consists of internal (or split) nodes and terminal (or leaf) nodes. Each split node will make a binary decision and the final decision is made based on the terminal node reached by the input feature. Tree-ensemble methods regard different decision trees as weak learners, and then construct a strong learner by either bagging or boosting. Bagging, also known as bootstrap aggregating (Breiman, 1996), is used to reduce the variance of the model. Multiple random subsets of the dataset with replacements are first selected, one for training an individual sub-model. Then an average prediction from these sub-models is calculated. Random Forest (Liaw and Wiener, 2002) extends the bagging by exploiting a small tweak that reduces the correlation between the bagged trees. For the boosting algorithm, the boosted tree (strong learner) is regarded as a combination of the single trees (weak learners). The weight of the combination is updated adaptively according to the different designs of the objective function and optimization methods. AdaBoost (Freund and Schapire, 1997) is the first version of the boosting method, in which the weak learners are iteratively trained on a weighted dataset by minimizing the exponential loss. XGBoost extends to more general loss function via gradient boosting optimization and learns a model with an additive training trick.

The objective of XGBoost is to learn a model with good variance-bias balance. In other words, the model should have strong predictive power but also large variance to be generalized on the extra data. This can be represented with the following objective function with respect to model parameter θ:

$$\mathrm{obj}(\theta) = \mathrm{L}(\theta) + \Omega(\theta)$$

where the first term is the loss function which should be minimized, and the second term is a regularization term of the model's complexity to prevent it from over-fitting. Considering a tree-ensemble model where the overall prediction is the summation of K predictive values across all the trees $\mathrm{f_k}(\mathrm{x_i})$,

$$p_i = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{f_k}(\mathrm{x_i}),$$

the objective function can be written as:

$$\mathrm{obj}(\theta) = \sum_{\mathrm{i}}^{\mathrm{n}} \mathrm{l}(\mathrm{p_i}, \mathrm{t_i}) + \sum_{\mathrm{k=1}}^{\mathrm{K}} \Omega(\mathrm{f_k}),$$

where $\mathrm{l}(\mathrm{p_i}, \mathrm{t_i})$ is the mean-squared loss imposed on each sample i regarding its predictive value $\mathrm{p_i}$ and the label $\mathrm{t_i}$, and $\Omega(\mathrm{f_k})$ is the regularization constraint imposed on each tree. XGBoost applies an efficient addictive training algorithm to optimize such an objective function. This algorithm will learn one tree at each step, then add a new tree by fixing what it has learned, mathematically,

$$\mathrm{p_i^{(0)}} = 0,$$

$$\mathrm{p_i^{(1)}} = \mathrm{f_1}(\mathrm{x_i}) = \mathrm{p_i^{(0)}} + \mathrm{f_1}(\mathrm{x_i}),$$

$$\cdots\cdots$$

$$\mathrm{p_i^{(t)}} = \sum_{k=1}^{t} \mathrm{f_k}(\mathrm{x_i}) = \mathrm{p_i^{(t-1)}} + \mathrm{f_t}(\mathrm{x_i}).$$

Thus, the objective at step t becomes,

$$\mathrm{obj}^{(t)} = \sum_{i=1}^{n} \left(t_i - \left(p_i^{(t-1)} + f_t(x_i)\right)\right)^2 + \sum_{i=1}^{t} \Omega(f_i)$$

$$= \sum_{i=1}^{n} \left[2\left(p_i^{(t-1)} - t_i\right) f_t(x_i) + f_t(x_i)^2\right] + \Omega(f_t) + \mathrm{constant}$$

This can be easily optimized with second-order Tylor expansion, considering the first and second-order gradients, $g_i = \partial_{p_i^{(t-1)}} l\left(t_i, p_i^{(t-1)}\right)$ and $h_i = \partial^2_{p_i^{(t-1)}} l\left(t_i, p_i^{(t-1)}\right)$ respectively, with the objective function at step t now becoming,

$$\mathrm{obj}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t)$$

To this end, we introduced how to efficiently train the boosted trees with an additive strategy, i.e. training a new tree at a step t by optimizing above step-based objective function. One of the merits of this definition is that the objective value only depends on the g_i and h_i, which allows using custom loss function. $\Omega(f_t)$ is the regularization term, which controls the complexity of the model. Now, we re-define the tree by a vector of prediction score in leaves,

$$f_t(x) = w_{q(x)}, w \in \mathbb{R}^T$$

where $q(x_i)$ is a mapping function that maps a training instance to a leaf. Based on this re-defined formulation, $\Omega(f)$ can be heuristically defined as

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2,$$

where T is the number of leaves of the tree and w_j is the prediction score in each leaf. By re-grouping the training samples on each leaf j, the objective function can hence be reformed as

$$obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2$$

$$= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T,$$

where $I_j = \{i|q(x_i) = j\}$ is the indices of training instances which reach the jth leaf. We use $G_j = \sum_{i\in I_j} g_i$ and $H_j = \sum_{i\in I_j} h_i$ to express the summation of first/second ordered gradients across leaves. Thus, the objective function can then be further simplified as

$$obj^{(\mathrm{t})} \approx \sum_{j=1}^{T}\left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right)w_j^2\right] + \gamma T$$

Note that w_j are independent with each other, thus the equation has a quadratic form, the solution for the above equation is

$$w_j{}^* = -\frac{G_j}{H_j + \lambda},$$

and the resulting objective value is

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j{}^2}{H_j + \lambda} + \gamma T.$$

This equation measures how good a tree structure $q(x_i)$ is for a certain training instance. Based on this property, one can grow a tree greedily using the information gain. To specify this information gain, we consider the gradients flow before and after splitting,

$$G_L = \sum_{j\in T_L} g_j\,, \qquad G_R = \sum_{j\in T_R} g_j$$

$$H_L = \sum_{j\in T_L} h_j\,, \qquad H_R = \sum_{j\in T_R} h_j$$

where T_L and T_R are the indices of left and right leaves respectively. Before splitting, the tree's complexity is

$$\frac{(G_L + G_R)^2}{H_L + H_R + \lambda} + \gamma.$$

After splitting, the tree has complexity,

$$\frac{G_L{}^2}{H_L + \lambda} + \frac{G_R{}^2}{H_R + \lambda} + 2\gamma,$$

Then the information gain of a splitting tree can be calculated as

$$Gain = \frac{1}{2}\left[\frac{{G_L}^2}{H_L + \lambda} + \frac{{G_R}^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma$$

As a result, we can outline the *XGBoost* algorithm as an iteration process. For each iteration, we perform the following operations: 1) Grow the tree to the maximum depth by finding the best splitting points via information gain. 2) Assign prediction score to the two new leaves. 3) Prune the tree by deleting the nodes with negative gain.

4 Implementation of XGBoost – a reliable prediction model

There exists some sparse data in our experimental dataset that needs the adoption of XGBoost. The dataset was arbitrarily split into two subsets; 75% of the data was used as a training set and 25% as a validation set. All the training data for Xbgoost was used to construct the model. The validation data was used to test the results with the data that was not utilized to develop the model. In order to improve the calculation efficiency, and prevent individual data from overflowing during the calculation, input and output parameters were normalized. In addition, all 14 variables show independence from each other after doing correlation analysis, which indicates these 14 variables can be used for predicting the environmental performance in a model.

PyCharm was adopted to train and develop the XGBoost model for corporate environmental performance. A statistical package scikit-learn in python was used to implement the XGBoost. To determine the hyper-parameters of the model, we applied a brute force grid search with 5-fold cross-validation. In order to achieve optimal parameter setting, we needed to initialize the searching with some prior knowledge of the parameters' ranges. For example, the learning rate for XGBoost is usually 0.05, and the maximum depth is usually 6, 7, or 8. Other parameters, such as 'min_child_weight', 'subsample', and 'colsample_bytree' need to be carefully tuned since they greatly affect the model's generalizability. Thus, we applied different seed during the searching to increase the variance of the model. The boosting iterations were determined using early stopping, and mean squared error was applied as the evaluation metrics during the searching. Table 4 shows the finally determined values for the hyper-parameters of the XGBoost model which achieve the best performance.

Table 4: Values Determined for the Hyper-parameters of the XGBoost Model

	Description	Value
'eta'	Boosting learning rate	0.03
'subsample'	Subsample ratio of the training instance	0.8
'colsample_bytree'	Subsample ratio of columns when constructing each tree	0.8
'objective'	Specify the learning task and the corresponding learning objective	'linear'
'max_depth'	Maximum tree depth for base learners	7
'min_child_weight'	Minimum sum of instance weight(hessian) needed in a child	0.5
'num_boost_round'	Number of boosting iterations	1000

4.1 Evaluation Criteria for model

The performance evaluation indices for the models tested in this paper are Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Correlation Coefficient (CC) and Coefficient of Determination (R-square, R^2), which are defined as follows:

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y_i}|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2}$$

$$\mathrm{CC} = \frac{\sum_{i=1}^{m}(x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x_i})^2 \sum_{i=1}^{m}(y_i - \overline{y_i})^2}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y_i})^2}{\sum_{i=1}^{m}(y_i - \overline{y_i})^2}$$

where y_i is the observed value for parameter y, $\hat{y_i}$ is the predicted value and $\overline{y_i}$ is the mean of observed values.

4.2 Reliability of XGBoost model

Random Forest (RF) and Support Vector Machines for regression (SVMreg) are commonly adopted machine learning methods when dealing with prediction problems (Chaudhuri and De, 2011; Lee, 2007; Pan, 2018; Tsanas and Xifara, 2012). In this research, RF and SVMreg prediction methods were implemented and compared with an XGBoost model.

Fig. 5 presents the initial data curve and relative error curve of the training set and testing data. For the training course curve and testing course curve, a dot was extracted from the curve every 10 samples, 88 samples in total. And for the training error and testing error curves, a dot was extracted from each curve every 3 samples, 73 samples each in total. It can be seen that the prediction relative errors of the training samples under the XGBoost model are nearly 0.04%, exhibiting much better performance compared to SVMReg and RF. This demonstrates that the developed XGBoost model can more precisely describe the complex relationship between corporate environmental performance and explanatory variables. The predicted environmental performance on validation data by the three models and the relative errors between the predicted value and real value are illustrated in Fig. 5(b), Fig. 5(d), and Fig. 5(f). The MAE, RMSE, CC, and R^2 of the testing samples under the three models are compared in Table 5.

Fig. 5 and Table 5 also show that using the XGBoost method to predict corporate environmental performance is better than using RF and SVMreg. The XGBoost method is more efficient and is a reliable alternative for corporate environmental performance prediction.

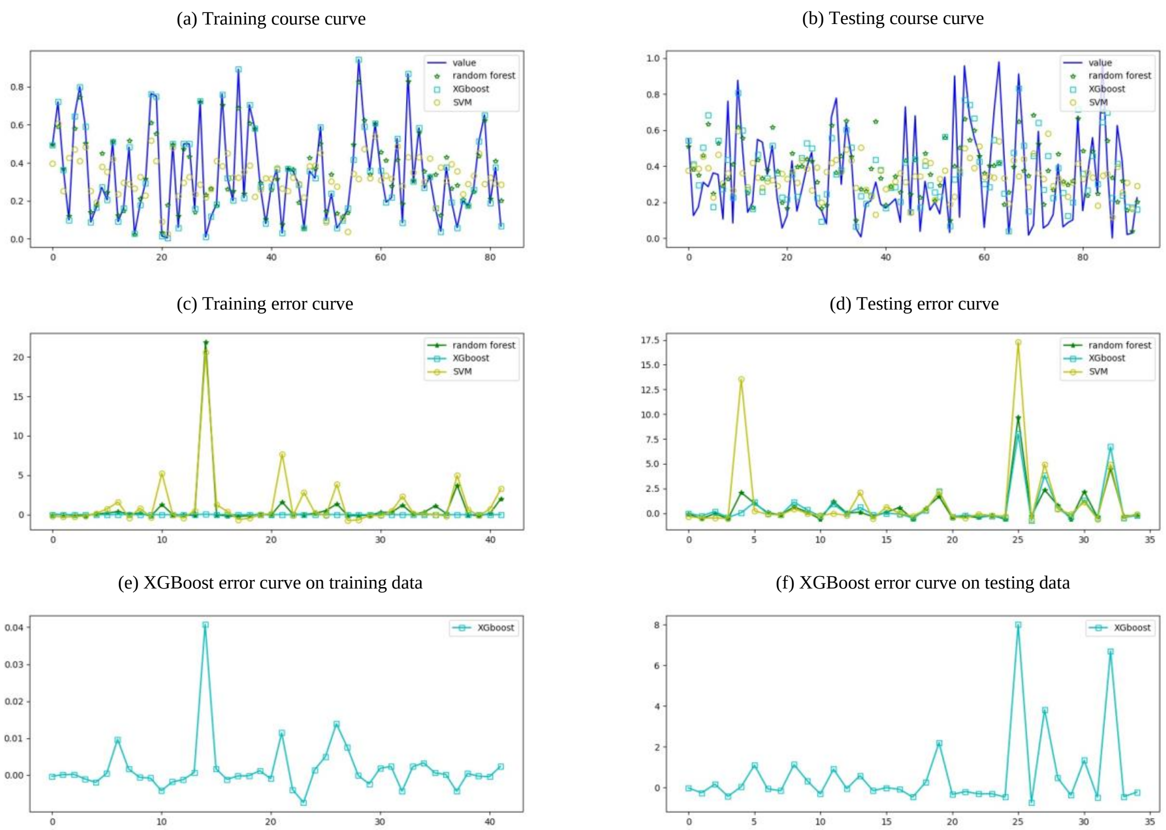

Figure 5: Course curves and relative error curves. (a) training course curve of initial data and three models, (b) testing course curve of initial data and three models, (c) training error curve of three models, (d) testing error curve of three models, (e) XGBoost error curve on training data, and (f) XGBoost error curve on testing data.

Table 5: Comparison of the prediction accuracy of SVMReg, RF, and XGBoost

Method	MAE	RMSE	CC	R2
SVMreg	0.19304	0.23527	0.39971	0.15951
RF	0.16578	0.20429	0.61295	0.36630
XGBoost	**0.14546**	**0.18336**	**0.70244**	**0.48952**

5 Empirical results and discussions

A SHAP value for a feature of a specific prediction represents how much the model prediction changes when we observe that feature. In the summary plot below (Fig. 6), all the SHAP values for a single feature on a row are drawn, where the x-axis is the SHAP value (which for this model is in units of log odds of corporate environmental performance).

Fig. 6 indicates that total asset (TA), amount of proactive environmental costs (PEC), proportion of technicians (PT) and number of patents (PA) were more important in this model while tax bracket (TB), formal legal warning since firm founding (FLW), equity (EQ), proportion of in-plant environment (PIE) and environmental projects payback period (PP) were relatively less important.

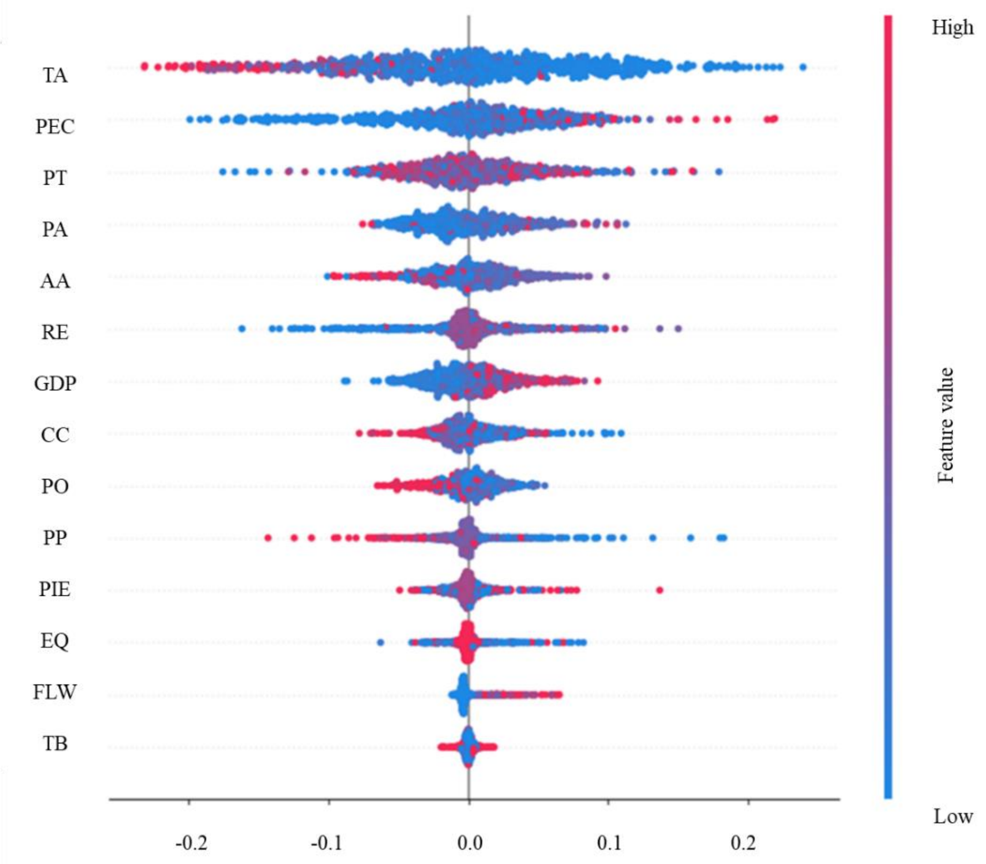

Figure 6: Summary of SHAP values for 14 variables (impact on model output)

This SHAP figure not only indicates which features are most important but also their range of effects over the dataset. Each dot is coloured by the value of that feature from high to low. For example, as shown in Fig. 6, red dots for the rows of ‘total asset’ tend to appear on the left side. This means that high values for total asset lead to lower corporate environmental performance, or in other words, total asset exhibit a positive relationship to corporate environmental performance. As shown in Fig. 6, GDP shows a positive relationship with corporate environmental performance, while population and consumption of coal have a negative effect on corporate environmental performance. Based on the analysis above, Hypothesis 1 cannot be supported. Amount of proactive environmental costs and number of formal legal warnings since firm founding reflect a positive relationship with corporate environmental performance. These results provide support for Hypothesis 2. It is revealed that the higher value for proportion of in-plant environment, the better environmental performance corporates will get. As such, Hypothesis 3 is supported. There indicates a negative effect on corporate environmental performance for total assets, environmental projects payback period and asset age. In contrast, a positive relationship exists between proportion of technicians, equity, revenue, and corporate environmental performance. Fewer total assets correlated with higher corporate environmental performance. As for tax bracket, the relationship reveals unclear. Thus, Hypothesis 4 cannot be supported. Fig. 6 clearly indicates that number of patents shows a positive relationship with corporate environmental performance, verifying Hypothesis 5.

Total assets (TA) represents the size and ability of a firm, which is highly related to corporate environmental performance (Trumpp Christoph and Guenther Thomas, 2017; Zhang et al., 2008). Amount of proactive environmental costs (PEC) is a direct reflection of investment in the environmental strategy of a firm. The proportion of technicians (PT) and number of patents (PA) show the technological innovation ability of a firm. Advanced technology can reduce the environmental impact of firms, improve energy efficiency, and increase corporate environmental performance (Dietz and Rosa, 1994; Wang et al., 2013). The variable explanation rankings show the firm characteristic variables explain more about the prediction model, indicating that Natural-resource-based view is better to study the corporate environmental performance. As researchers pointed out before, the external environment, including normative and regulative environmental, remains undeveloped and fragmented in China (Gao et al., 2019). Thus, the Institutional theory explained less about corporate environmental performance in China.

There are massive reasons leading to the complex relationship between variables and corporate environmental performance. Amount of proactive environmental costs is the direct investment in environmental strategy, and the result is similar with the previous research which indicates that greater investment leads to higher corporate environmental performance (Fu et al., 2017). The negative relationship between formal legal warnings since firm founding could due to that the corporates need to keep its positive image. If a firm receives a legal warning that damages its social image, it may adopt measures to mitigate this effect, such as developing and implementing environmental strategies.

The negative relationship between total assets and corporate environmental performance stands in contrast to previous research findings (Al-Tuwaijri et al., 2004). This may be because the proportion of assets dedicated to environmental investment by large firms is relatively low although the large firms care more about social responsibility (Udayasankar, 2008). Firms survive based on their profitability, which enforces firms to invest in profitable projects. As for the payback period, firms tend to invest in projects with short payback period to avoid the risks. If the payback period is too long, firms will engage in these projects, leading to less environmental improvement projects and poorer environmental performance. Revenue can show the profitability of a firm and firms with strong profitability tend to pay more attention to environmental issues (Orlitzky et al., 2003; Russo and Fouts, 1997). Regions with higher GDP are usually developed regions in China, such as Beijing, Shanghai, Jiangsu, and Guangdong. These regions are stricter about environmental protection and have inaugurated several policies regarding environmental sustainability (Zheng et al., 2018).

Based on the findings of positive and negative relationships between variables and corporate environmental performance, several policies can be put in place to improve corporate environmental performance. To increase firms' contributions to corporate proactive environmental costs, the government should provide more financial incentives for environmental protection, including tax benefits, green loans, and environmental subsidies. When providing these incentives, the payback period should be taken into consideration since longer periods entail more risks. Currently many corporations have spent hundreds of millions of dollars on environmental projects (Berry and Rondinelli, 1998). Fines for noncompliance need to be increased and enforcement of environmental regulations should be strengthened, making business executives and owners liable for environmental pollution. The number of fines and intensity of enforcement also need to be applied in accordance with the size of the corporation. Corporate environmental issues need to receive more attention in regions with lower GDP. The local governments of these regions can learn from the experiences of more developed regions. At present, the incentives are primarily provided to the larger-scale corporations since they demonstrate better financial performance. However, as knowledge, practices, systems, and routines at the business and natural environment interface become more widely dispersed, smaller companies may also begin to adopt voluntary niche environmental strategies (Sharma, 2000). Governments ought to promote the environmental strategies of small companies and develop some targeted incentives tailored to them.

Whatever policies the government may implement, corporations could internally choose to direct more investment toward environmental prevention and minimization. Introducing advanced technologies, employing more technicians, reusing materials, and adopting an environmental corporate culture are other advisable measures. Full-cost accounting is suggested for adoption when managing the corporations, considering direct costs (labour, capital, and raw materials), hidden costs (monitoring and reporting;), contingent liability costs (fines and remedial action), and less tangible costs (public relations and goodwill). Corporations can use full-cost accounting to choose the most eco-effective projects and improve corporate environmental performance.

6 Conclusion

In this study, we identified the relationship between institutional context, corporate environmental performance with corporate environmental performance based on a combination of Institutional Theory and the Natural-Resource-Based View. We presented an approach to conducting this identification by predicting corporate environmental performance with machine learning methods. The key challenge of dealing with noise in the data from ESCOs that compromises the accuracy of these predictive models was also highlighted. In this study, a newly introduced machine learning algorithm, XGBoost, was applied to predict corporate environmental performance. Data from 1100 projects for ESCOs in the time period between 2011 and 2015 was analyzed to explore the statistical relationship between 14 input variables (GDP, population, consumption of coal, amount of proactive environmental costs, number of formal legal warnings since a firm's founding, proportion of in-plant area, proportion of technicians, total worth, equity, environmental projects payback period, asset age, revenue, tax bracket, number of patents) and the output variable, corporate environmental performance. The results indicate that XGBoost achieved higher accuracy than other learning algorithms and was reliable to test the relationship.

The findings of this research agree with those in the machine learning literature strongly endorsing the use of XGBoost in complex applications (Gumus and Kiran, 2017; Pan, 2018). Applying SHAP in XGBoost model interpretation enables the impact of input variables on the output to be determined. In the model, total assets (TA), amount of proactive environmental costs (PEC), proportion of technicians (PT) and number of patents (PA) are found to contribute the most to corporate environmental performance. Also, the impacts each feature has on the model output was obtained through SHAP summary plotting. Amount of proactive environmental costs (PEC), Revenue (RE), GDP, and number of formal legal warnings since the firm's founding (FLW) show a positive relationship with corporate environmental performance, while total assets (TA) and environmental projects payback period (PP) show a negative relationship. Based on the SHAP findings, several policy recommendations and environmental strategies for governments and corporations to carry out are proposed to improve corporate environmental performance. Corporates with stronger governance capability, information management capability and technology-related capability will perform better corporate environmental performance.

Although this paper contributes to corporate environmental performance, there are still some research limitations. First, the prediction accuracy for all observations is relatively low due to the result of noisy data and the limited input gaps between machine learning and social sciences (CHEN et al., 2018). The rate may increase if more information about corporate environmental performance is considered. Second, the data used are only from the ESCO industry in China. It could add more value if the model can be tested in other industries and in other countries.

In future studies, more variables and more data should be introduced to achieve greater accuracy in predicting corporate environmental performance. Since corporates in China are becoming increasingly aware of the environmental performance, along with increasing national policies regarding corporate environmental performance, more variables from the perspective of institutional theory can be taken into consideration. In addition, due to the contrast results with previous research about the relationship between total assets and corporate environmental performance, the total assets could be considered to have a nonlinear relationship as the moderators when investigating the relationship between corporate financial performance and corporate environmental performance.

References

Ajit, P., Punnoose R, 2016. Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. Int. J. Adv. Res. Artif. Intell. 5, 22–26.

Aldrich, H.E., Fiol, C.M., 1994. Fools Rush in? The Institutional Context of Industry Creation. Acad. Manage. Rev. 19, 645–670. https://doi.org/10.5465/AMR.1994.9412190214

Al-Tuwaijri, S.A., Christensen, T.E., Hughes, K.E., 2004. The relations among environmental disclosure, environmental performance, and economic performance: a simultaneous equations approach. Account. Organ. Soc. 29, 447–471. https://doi.org/10.1016/S0361-3682(03)00032-1

Backman, C.A., Verbeke, A., Schulz, R.A., 2017. The Drivers of Corporate Climate Change Strategies and Public Policy: A New Resource-Based View Perspective. Bus. Soc. 56, 545–575. https://doi.org/10.1177/0007650315578450

Bansal, P., Clelland, I., 2004. Talking Trash: Legitimacy, Impression Management, and Unsystematic Risk in the Context of the Natural Environment. Acad. Manage. J. 47, 93–103. https://doi.org/10.2307/20159562

Bansal, P., Gao, J., 2006. Building the Future by Looking to the Past: Examining Research Published on Organizations and Environment. Organ. Environ. 19, 458–478. https://doi.org/10.1177/1086026606294957

Bansal Pratima, 2005. Evolving sustainably: a longitudinal study of corporate sustainable development. Strateg. Manag. J. 26, 197–218. https://doi.org/10.1002/smj.441

Benitez-Amado, J., Walczuch, R.M., 2012. Information technology, the organizational capability of proactive corporate environmental strategy and firm performance: a resource-based analysis. Eur. J. Inf. Syst. 21, 664–679. https://doi.org/10.1057/ejis.2012.14

Berrone, P., Gomez-Mejia, L.R., 2009. Environmental Performance and Executive Compensation: An Integrated Agency-Institutional Perspective. Acad. Manage. J. 52, 103–126.

Berry, M.A., Rondinelli, D.A., 1998. Proactive Corporate Environmental Management: A New Industrial Revolution. Acad. Manag. Exec. 1993-2005 12, 38–50.

Bhatnagar, V., 1999. Evaluating corporate environmental performance in developing countries. Sustain. Meas. Eval. Report. Environ. Soc. Perform. Greenleaf Publ. Shefffield S3 8GG UK.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Buysse, K., Verbeke, A., 2003. Proactive environmental strategies: a stakeholder management perspective. Strateg. Manag. J. 24, 453–470. https://doi.org/10.1002/smj.299

Camisón, C., 2010. Effects of coercive regulation versus voluntary and cooperative auto-regulation on environmental adaptation and performance: Empirical evidence in Spain. Eur. Manag. J. 28, 346–361. https://doi.org/10.1016/j.emj.2010.03.001

Campbell, J.L., 2007. Why Would Corporations Behave in Socially Responsible Ways? An Institutional Theory of Corporate Social Responsibility. Acad. Manage. Rev. 32, 946–967. https://doi.org/10.2307/20159343

Chan, C.M., Makino, S., 2007. Legitimacy and multi-level institutional environments: implications for foreign subsidiary ownership structure. J. Int. Bus. Stud. 38, 621–638. https://doi.org/10.1057/palgrave.jibs.8400283

Chang, L., Li, W., Lu, X., 2015. Government Engagement, Environmental Policy, and Environmental Performance: Evidence from the Most Polluting Chinese Listed Firms. Bus. Strategy Environ. 24, 1–19. https://doi.org/10.1002/bse.1802

Chaudhuri, A., De, K., 2011. Fuzzy Support Vector Machine for bankruptcy prediction. Appl. Soft Comput., The Impact of Soft Computing for the Progress of Artificial Intelligence 11, 2472–2486. https://doi.org/10.1016/j.asoc.2010.10.003

CHEN, N.-C., DROUHARD, M., KOCIELNIK, R., SUH, J., ARAGON, C.R., 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting The Focus to Ambiguity.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16. ACM, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Christmann, P., 2004. Multinational Companies and the Natural Environment: Determinants of Global Environmental Policy. Acad. Manage. J. 47, 747–760. https://doi.org/10.2307/20159616

Colwell, S.R., Joshi, A.W., 2013. Corporate Ecological Responsiveness: Antecedent Effects of Institutional Pressure and Top Management Commitment and Their Impact on Organizational Performance. Bus. Strategy Environ. 22, 73–91. https://doi.org/10.1002/bse.732

Crowley, K., 2013. Irresistible force? Achieving carbon pricing in Australia. Aust. J. Polit. Hist. 59, 368–381.

Cui, L., Jiang, F., 2012. State ownership effect on firms’ FDI ownership decisions under institutional pressure: a study of Chinese outward-investing firms. J. Int. Bus. Stud. 43, 264–284. https://doi.org/10.1057/jibs.2012.1

Delmas, M., Blass, V.D., 2010. Measuring corporate environmental performance: the trade-offs of sustainability ratings. Bus. Strategy Environ. 19, 245–260. https://doi.org/10.1002/bse.676

Delmas, M., Toffel, M.W., 2004. Stakeholders and environmental management practices: an institutional framework. Bus. Strategy Environ. 13, 209–222. https://doi.org/10.1002/bse.409

Deng, X., Zheng, S., Xu, P., Zhang, X., 2017. Study on dissipative structure of China’s building energy service industry system based on brusselator model. J. Clean. Prod. 150, 112–122. https://doi.org/10.1016/j.jclepro.2017.02.198

Dibrell, C., Craig, J., Hansen, E., 2011. Natural Environment, Market Orientation, and Firm Innovativeness: An Organizational Life Cycle Perspective. J. Small Bus. Manag. 49, 467–489. https://doi.org/10.1111/j.1540-627X.2011.00333.x

Dietz, T., Rosa, E.A., 1994. Rethinking the Environmental Impacts of Population, Affluence and Technology. Hum. Ecol. Rev. 1, 277–300.

DiMaggio, P.J., Powell, W.W., 2000. The iron cage revisited institutional isomorphism and collective rationality in organizational fields, in: Economics Meets Sociology in Strategic Management. Emerald Group Publishing Limited, pp. 143–166.

Dixon-Fowler, H.R., Slater, D.J., Johnson, J.L., Ellstrand, A.E., Romi, A.M., 2013. Beyond "Does it Pay to be Green?" A Meta-Analysis of Moderators of the CEP—CFP Relationship. J. Bus. Ethics 112, 353–366.

Etzion, D., 2007. Research on Organizations and the Natural Environment, 1992-Present: A Review. J. Manag. 33, 637–664. https://doi.org/10.1177/0149206307302553

Freeman, R.E., 1984. Strategic management: a stakeholder approach. Pitman, Boston.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

Friedman, M., 2007. The Social Responsibility of Business Is to Increase Its Profits, in: Zimmerli, W.C., Holzinger, M., Richter, K. (Eds.), Corporate Ethics and Corporate Governance. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 173–178. https://doi.org/10.1007/978-3-540-70818-6_14

Fu, X., Lahr, M., Yaxiong, Z., Meng, B., 2017. Actions on climate change, Intended Reducing carbon emissions in China via optimal industry shifts: Toward hi-tech industries, cleaner resources and higher carbon shares in less-develop regions. Energy Policy. https://doi.org/10.1016/j.enpol.2016.10.038

Gallego-Alvarez, I., Ortas, E., Vicente-Villardón, J.L., Etxeberria, I.Á., 2017. Institutional Constraints, Stakeholder Pressure and Corporate Environmental Reporting Policies. Bus. Strategy Environ. 26, 807–825. https://doi.org/10.1002/bse.1952

Gangi, F., Mustilli, M., Varrone, N., 2019. The impact of corporate social responsibility (CSR) knowledge on corporate financial performance: evidence from the European banking industry. J. Knowl. Manag. https://doi.org/10.1108/JKM-04-2018-0267

Gao, L.S., Connors, E., 2011. Corporate Environmental Performance, Disclosure and Leverage: An Integrated Approach. Int. Rev. Account. Bank. Finance 1.

Gao, Y., Gu, J., Liu, H., 2019. Interactive effects of various institutional pressures on corporate environmental responsibility: Institutional theory and multilevel analysis. Bus. Strategy Environ. 28, 724–736. https://doi.org/10.1002/bse.2276

Garnaut, R., 2008. The Garnaut climate change review. Camb. Camb.

Gumus, M., Kiran, M.S., 2017. Crude oil price forecasting using XGBoost, in: 2017 International Conference on Computer Science and Engineering (UBMK). Presented at the 2017 International

Conference on Computer Science and Engineering (UBMK), pp. 1100–1103. https://doi.org/10.1109/UBMK.2017.8093500

Hall, J., Wagner, M., 2012. Integrating Sustainability into Firms’ Processes: Performance Effects and the Moderating Role of Business Models and Innovation. Bus. Strategy Environ. 21, 183–196. https://doi.org/10.1002/bse.728

Hart, S.L., 2010. Capitalism at the crossroads: Next generation business strategies for a post-crisis world. FT Press.

Hart, S.L., 1995. A Natural-Resource-Based View of the Firm. Acad. Manage. Rev. 20, 986–1014. https://doi.org/10.5465/AMR.1995.9512280033

Hoi, C.K., Wu, Q., Zhang, H., 2013. Is Corporate Social Responsibility (CSR) Associated with Tax Avoidance? Evidence from Irresponsible CSR Activities. Account. Rev. 88, 2025–2059. https://doi.org/10.2308/accr-50544

Ilinitch, A.Y., Soderstrom, N.S., E. Thomas, T., 1998. Measuring corporate environmental performance. J. Account. Public Policy 17, 383–408. https://doi.org/10.1016/S0278-4254(98)10012-1

IPCC, 2015. Climate change 2014: mitigation of climate change. Cambridge University Press.

Jennings, P.D., Zandbergen, P.A., 1995. Ecologically Sustainable Organizations: An Institutional Approach. Acad. Manage. Rev. 20, 1015–1052. https://doi.org/10.2307/258964

Kagan, R.A., Gunningham, N., Thornton, D., 2003. Explaining Corporate Environmental Performance: How Does Regulation Matter? Law Soc. Rev. 37, 51–90. https://doi.org/10.1111/1540-5893.3701002

King, A.A., Lenox, M.J., 2001. Does It Really Pay to Be Green? An Empirical Study of Firm Environmental and Financial Performance: An Empirical Study of Firm Environmental and Financial Performance. J. Ind. Ecol. 5, 105–116. https://doi.org/10.1162/108819801753358526

Kitada, M., Ölçer, A., 2015. Managing people and technology: The challenges in CSR and energy efficient shipping. Res. Transp. Bus. Manag., Energy Efficiency in Maritime Logistics Chains 17, 36–40. https://doi.org/10.1016/j.rtbm.2015.10.002

Klassen, R.D., McLaughlin, C.P., 1996. The Impact of Environmental Management on Firm Performance. Manag. Sci. 42, 1199–1214. https://doi.org/10.1287/mnsc.42.8.1199

Lee, J.W., Kim, Y.M., Kim, Y.E., 2018. Antecedents of Adopting Corporate Environmental Responsibility and Green Practices. J. Bus. Ethics 148, 397–409. https://doi.org/10.1007/s10551-016-3024-y

Lee, Y.-C., 2007. Application of support vector machines to corporate credit rating prediction. Expert Syst. Appl. 33, 67–74. https://doi.org/10.1016/j.eswa.2006.04.018

Li, D., Zhao, Y., Sun, Y., Yin, D., 2017. Corporate environmental performance, environmental information disclosure, and financial performance: Evidence from China. Hum. Ecol. Risk Assess. Int. J. 23, 323–339. https://doi.org/10.1080/10807039.2016.1247256

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Liu, G., Zheng, S., Xu, P., Zhuang, T., 2018. An ANP-SWOT approach for ESCOs industry strategies in Chinese building sectors. Renew. Sustain. Energy Rev. 93, 90–99. https://doi.org/10.1016/j.rser.2018.03.090

Liu, Z., Davis, S.J., Feng, K., Hubacek, K., Liang, S., Anadon, L.D., Chen, B., Liu, J., Yan, J., Guan, D., 2016. Targeted opportunities to address the climate–trade dilemma in China. Nat. Clim. Change 6, 201. https://doi.org/10.1038/nclimate2800

Lober, D.J., 1996. Evaluating The Environmental Performance Of Corporations. J. Manag. Issues 8, 184–205.

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774.

Luxmore, S.R., Hull, C.E., Tang, Z., 2018. Institutional Determinants of Environmental Corporate Social Responsibility: Are Multinational Entities Taking Advantage of Weak Environmental Enforcement in Lower-Income Nations? Bus. Soc. Rev. 123, 151–179. https://doi.org/10.1111/basr.12138

Lyon, T.P., Maxwell, J.W., 1999. Corporate environmental strategies as tools to influence regulation. Bus. Strategy Environ. 8, 189–196. https://doi.org/10.1002/(SICI)1099-0836(199905/06)8:3<189::AID-BSE194>3.0.CO;2-0

McWilliams, A., Siegel, D., 2001. Corporate social responsibility: A theory of the firm perspective. Acad. Manage. Rev. 26, 117–127.

McWilliams, A., Siegel, D.S., Wright, P.M., 2006. Corporate Social Responsibility: Strategic Implications*. J. Manag. Stud. 43, 1–18. https://doi.org/10.1111/j.1467-6486.2006.00580.x

Meng, X.H., Zeng, S.X., Shi, J.J., Qi, G.Y., Zhang, Z.B., 2014. The relationship between corporate environmental performance and environmental disclosure: An empirical study in China. J. Environ. Manage. 145, 357–367. https://doi.org/10.1016/j.jenvman.2014.07.009

Meyer, J.W., Rowan, B., 1977. Institutionalized Organizations: Formal Structure as Myth and Ceremony. Am. J. Sociol. 83, 340–363.

Millar, R.J., Hepburn, C., Beddington, J., Allen, M.R., 2018. Principles to guide investment towards a stable climate. Nat. Clim. Change 8, 2. https://doi.org/10.1038/s41558-017-0042-4

Muller, A., Kolk, A., 2010. Extrinsic and Intrinsic Drivers of Corporate Social Performance: Evidence from Foreign and Domestic Firms in Mexico. J. Manag. Stud. 47, 1–26. https://doi.org/10.1111/j.1467-6486.2009.00855.x

Neuhoff, K., 2011. Climate policy after Copenhagen: the role of carbon pricing. Cambridge University Press.

Newell, P., Paterson, M., 2010. Climate capitalism: global warming and the transformation of the global economy. Cambridge University Press.

Nyberg, D., Spicer, A., Wright, C., 2013. Incorporating citizens: corporate political engagement with climate change in Australia. Organization 20, 433–453.

Orlitzky, M., Schmidt, F.L., Rynes, S.L., 2003. Corporate Social and Financial Performance: A Meta-Analysis. Organ. Stud. 24, 403–441. https://doi.org/10.1177/0170840603024003910

Pan, B., 2018. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. IOP Conf. Ser. Earth Environ. Sci. 113, 012127. https://doi.org/10.1088/1755-1315/113/1/012127

Post, C., Rahman, N., McQuillen, C., 2015. From Board Composition to Corporate Environmental Performance Through Sustainability-Themed Alliances. J. Bus. Ethics 130, 423–435. https://doi.org/10.1007/s10551-014-2231-7

Qiu, J., 2011. China to Spend Billions Cleaning Up Groundwater. Science 334, 745–745. https://doi.org/10.1126/science.334.6057.745

Renwick, D.W., Redman, T., Maguire, S., 2013. Green human resource management: A review and research agenda. Int. J. Manag. Rev. 15, 1–14.

Rockness, J.W., 1985. An Assessment of the Relationship Between Us Corporate Environmental Performance and Disclosure. J. Bus. Finance Account. 12, 339–354. https://doi.org/10.1111/j.1468-5957.1985.tb00838.x

Russo, M.V., Fouts, P.A., 1997. A Resource-Based Perspective On Corporate Environmental Performance And Profitability. Acad. Manage. J. 40, 534–559. https://doi.org/10.2307/257052

Salo, J., 2008. Corporate Governance and Environmental Performance: Industry and Country Effects. Compet. Change 12, 328–354. https://doi.org/10.1179/102452908X357293

Scott, W.R., 2013. Institutions and organizations: Ideas, interests, and identities. Sage Publications.

Sharfman, M.P., Shaft, T.M., Tihanyi, L., 2004. A Model of the Global and Institutional Antecedents of High-Level Corporate Environmental Performance. Bus. Soc. 43, 6–36. https://doi.org/10.1177/0007650304262962

Sharma, S., 2000. Managerial Interpretations and Organizational Context as Predictors of Corporate Choice of Environmental Strategy. Acad. Manage. J. 43, 681–697. https://doi.org/10.5465/1556361

Shrivastava, P., 1995. Environmental technologies and competitive advantage. Strateg. Manag. J. 16, 183–200. https://doi.org/10.1002/smj.4250160923

Spicer, B.H., 1978. Investors, Corporate Social Performance and Information Disclosure: An Empirical Study. Account. Rev. 53, 94–111.

Stanwick, P.A., Stanwick, S.D., 1998. The Relationship between Corporate Social Performance, and Organizational Size, Financial Performance, and Environmental Performance: An Empirical Examination. J. Bus. Ethics 17, 195–204.

Stern, N., 2008. The economics of climate change. Am. Econ. Rev. 98, 1–37.

Tashman, P., Rivera, J., 2016. Ecological uncertainty, adaptation, and mitigation in the U.S. ski resort industry: Managing resource dependence and institutional pressures. Strateg. Manag. J. 37, 1507–1525. https://doi.org/10.1002/smj.2384

Torlay, L., Perrone-Bertolotti, M., Thomas, E., Baciu, M., 2017. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform. 4, 159–169. https://doi.org/10.1007/s40708-017-0065-7

Trumpp, C., Endrikat, J., Zopf, C., Guenther, E., 2015. Definition, Conceptualization, and Measurement of Corporate Environmental Performance: A Critical Examination of a Multidimensional Construct. J. Bus. Ethics 126, 185–204. https://doi.org/10.1007/s10551-013-1931-8

Trumpp Christoph, Guenther Thomas, 2017. Too Little or too much? Exploring U-shaped Relationships between Corporate Environmental Performance and Corporate Financial Performance. Bus. Strategy Environ. 26, 49–68. https://doi.org/10.1002/bse.1900

Tsanas, A., Xifara, A., 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy Build. 49, 560–567. https://doi.org/10.1016/j.enbuild.2012.03.003

Tso, G.K.F., Yau, K.K.W., 2007. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy 32, 1761–1768. https://doi.org/10.1016/j.energy.2006.11.010

Tyteca, D., Carlens, J., Berkhout, F., Hertin, J., Wehrmeyer, W., Wagner, M., 2002. Corporate environmental performance evaluation: evidence from the MEPI project. Bus. Strategy Environ. 11, 1–13. https://doi.org/10.1002/bse.312

Udayasankar, K., 2008. Corporate Social Responsibility and Firm Size. J. Bus. Ethics 83, 167–175. https://doi.org/10.1007/s10551-007-9609-8

Veleva, V., Ellenbecker, M., 2001. Indicators of sustainable production: framework and methodology. J. Clean. Prod. 9, 519–549. https://doi.org/10.1016/S0959-6526(01)00010-5

Walls, J.L., Berrone, P., Phan, P.H., 2012. Corporate governance and environmental performance: is there really a link? Strateg. Manag. J. 33, 885–913. https://doi.org/10.1002/smj.1952

Wang, H., Bi, J., Wheeler, D., Wang, J., Cao, D., Lu, G., Wang, Y., 2004. Environmental performance rating and disclosure: China's GreenWatch program. J. Environ. Manage. 71, 123–133. https://doi.org/10.1016/j.jenvman.2004.01.007

Wang, Q., Zhao, Z., Zhou, P., Zhou, D., 2013. Energy efficiency and production technology heterogeneity in China: A meta-frontier DEA approach. Econ. Model. 35, 283–289. https://doi.org/10.1016/j.econmod.2013.07.017

Wang, S., Li, J., Zhao, D., 2018. Institutional Pressures and Environmental Management Practices: The Moderating Effects of Environmental Commitment and Resource Availability. Bus. Strategy Environ. 27, 52–69. https://doi.org/10.1002/bse.1983

Wiedenhofer, D., Guan, D., Liu, Z., Meng, J., Zhang, N., Wei, Y.-M., 2017. Unequal household carbon footprints in China. Nat. Clim. Change 7, 75–80. https://doi.org/10.1038/nclimate3165

Xu, P., Chan, E.H.W., 2013. ANP model for sustainable Building Energy Efficiency Retrofit (BEER) using Energy Performance Contracting (EPC) for hotel buildings in China. Habitat Int., Low-Carbon Cities and Institutional Response 37, 104–112. https://doi.org/10.1016/j.habitatint.2011.12.004

Xu, P., Chan, E.H.W., Visscher, H.J., Zhang, X., Wu, Z., 2015. Sustainable building energy efficiency retrofit for hotel buildings using EPC mechanism in China: analytic Network Process (ANP) approach. J. Clean. Prod. 107, 378–388. https://doi.org/10.1016/j.jclepro.2014.12.101

Yeeles, A., 2018. Business as usual. Nat. Clim. Change 8, 13. https://doi.org/10.1038/s41558-017-0051-3

Yoon, Y., Gürhan-Canli, Z., Schwarz, N., 2006. The Effect of Corporate Social Responsibility (CSR) Activities on Companies With Bad Reputations. J. Consum. Psychol. 16, 377–390. https://doi.org/10.1207/s15327663jcp1604_9

Zhang, B., Bi, J., Yuan, Z., Ge, J., Liu, B., Bu, M., 2008. Why do firms engage in environmental management? An empirical study in China. J. Clean. Prod. 16, 1036–1045. https://doi.org/10.1016/j.jclepro.2007.06.016

Zhang, K., Wen, Z., Peng, L., 2007. Environmental Policies in China: Evolvement, Features and Evaluation. China Popul. Resour. Environ. 17, 1–7. https://doi.org/10.1016/S1872-583X(07)60006-0

Zheng, S., Alvarado, V., Xu, P., Leu, S.-Y., Hsu, S.-C., 2018. Exploring spatial patterns of carbon dioxide emission abatement via energy service companies in China. Resour. Conserv. Recycl. 137, 145–155. https://doi.org/10.1016/j.resconrec.2018.06.004

Appendix-XGBoost algorithm

```python
from XGBoost import plot_tree
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from pandas import read_csv, read_excel
import XGBoost as xgb
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error,r2_score, mean_absolute_error
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import Imputer, StandardScaler
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.base import BaseEstimator, TransformerMixin
import matplotlib.pyplot as plt
from sklearn import svm
import shap

data= pd.read_excel("data/111.xlsx")
data = data.drop(['NO.'],axis=1)
label = data.pop('Y2')



def main():
    # split data into train and test sets
    seed = 7
    test_size = .25
```

```
    X_train, X_test, y_train, y_test = train_test_split(data, label, test_size=test_size,
random_state=seed)
    original_col = X_train.columns
    imp = Imputer(missing_values='NaN', strategy='mean', axis=0)
    imp.fit(X_train)
    X_train = imp.transform(X_train)
    X_test = imp.transform(X_test)
    # random forest algorithm
    regr_rf = RandomForestRegressor(max_depth=30, random_state=2)
    regr_rf.fit(X_train, y_train)
    y_pred_train1= regr_rf.predict(X_train)
    y_pred1 = regr_rf.predict(X_test)
    # random forest end
    # XGBoost algorithm
    xgdmat=xgb.DMatrix(X_train,y_train)
    our_params={'eta':.03,'seed':0,'subsample':0.8,\
            'colsample_bytree':0.8,'objective':'reg:linear',\
            'max_depth':7,'min_child_weight':.5}
    # train the model
    final_gb=xgb.train(our_params,xgdmat,num_boost_round=1500)
    testmat = xgb.DMatrix(X_test)
    trainmat=xgb.DMatrix(X_train)
    y_pred2 = final_gb.predict(testmat)
    y_pred_train2= final_gb.predict(trainmat)
    # XGBoost end
    # svm regression
    clf = svm.SVR(kernel='rbf', degree = 3, gamma = 'auto', coef0=0.0, tol=0.1, C=1.0, epsilon=0.1,
shrinking = True, cache_size=200, verbose=False, max_iter=-1)
    clf.fit(X_train, y_train)
    y_pred_train3 = clf.predict(X_train)
    y_pred3 = clf.predict(X_test)
    # end svm
#random forest
    mae = mean_absolute_error(y_test.values, y_pred1)
    print("MAE: %.5f" % mae)
    rmse =np.sqrt(mean_squared_error(y_test.values, y_pred1))
    print("RMSE: %.5f" % rmse)
    R = np.corrcoef(y_test.values,y_pred1)
    print("Correlation Coef: %.5f" % R[0,1])
    r2 = r2_score(y_test.values,y_pred1)
    print("r2 score: %.5f" % r2)
 #XGBoost
    mae = mean_absolute_error(y_test.values, y_pred2)
    print("MAE: %.5f" % mae)
    rmse =np.sqrt(mean_squared_error(y_test.values, y_pred2))
```

```
    print("RMSE: %.5f" % rmse)
    R = np.corrcoef(y_test.values,y_pred2)

    print("Correlation Coef: %.5f" % R[0,1])
    r2 = r2_score(y_test.values,y_pred2)
    print("r2 score: %.5f" % r2)

    #svm
    mae = mean_absolute_error(y_test.values, y_pred3)
    print("MAE: %.5f" % mae)
    rmse =np.sqrt(mean_squared_error(y_test.values, y_pred3))
    print("RMSE: %.5f" % rmse)
    R = np.corrcoef(y_test.values,y_pred3)

    print("Correlation Coef: %.5f" % R[0,1])
    r2 = r2_score(y_test.values,y_pred3)
    print("r2 score: %.5f" % r2)

# #plot predict error
    plt.gcf().set_size_inches((10, 4))

    plt.plot(((y_pred1-y_test.values)/y_test.values)[::8], color='g', marker='*', label='random forest')
    plt.plot(((y_pred2-y_test.values)/y_test.values)[::8], color='c', marker='s', markerfacecolor='none',
label='XGBoost')
    plt.plot(((y_pred3-y_test.values)/y_test.values)[::8], color='y', marker='o',
markerfacecolor='none', label='SVM')
    # plt.gca().legend()
    plt.legend(loc='upper right')
    plt.savefig('junk.jpg')

# plot training error
    plt.gcf().set_size_inches((10, 4))
    plt.plot(((y_pred_train1-y_train.values)/y_train.values)[::20], color='g', marker='*',
label='random forest')
    plt.plot(((y_pred_train2-y_train.values)/y_train.values)[::20], color='c', marker='s',
markerfacecolor='none', label='XGBoost')
    plt.plot(((y_pred_train3-y_train.values)/y_train.values)[::20],color='y', marker='o',
markerfacecolor='none', label='SVM')
    # plt.gca().legend()
    plt.legend(loc='upper right')
    plt.savefig('junk.jpg')


# plot predict test
    plt.gcf().set_size_inches((10, 4))
    plt.plot(y_test.values[::3], color='b', label='value')
    plt.plot(y_pred1[::3], color='g', marker='*', markerfacecolor='none', label='random
forest',linestyle='None')
    plt.plot(y_pred2[::3], color='c', marker='s', markerfacecolor='none',
label='XGBoost',linestyle='None')
    plt.plot(y_pred3[::3], color='y', marker='o', markerfacecolor='none',
label='SVM',linestyle='None')
    # plt.gca().legend()
    plt.legend(loc='upper right')
    plt.savefig('junk.jpg')
```

```
#plot training data
    plt.gcf().set_size_inches((10, 4))
    plt.plot(y_train.values[::10], color='b', label='value')
    plt.plot(y_pred_train1[::10], color='g', marker='*', markerfacecolor='none', label='random forest',linestyle='None')
    plt.plot(y_pred_train2[::10], color='c', marker='s', markerfacecolor='none', label='XGBoost',linestyle='None')
    plt.plot(y_pred_train3[::10], color='y', marker='o', markerfacecolor='none', label='SVM',linestyle='None')
    # plt.gca().legend()
    plt.legend(loc='upper right')
    plt.savefig('junk2.jpg')

# shap value
    shap.initjs()
    shap_values = shap.TreeExplainer(final_gb).shap_values(X_train)
    X_train = pd.DataFrame(data=X_train, columns=original_col)
    X_train = X_train.rename(columns={
        "X2": "X7", "X3": "X6","X4":"X14","X5":"X4","X6":"X8","X7":"X9","X8":"X10","X9":"X12",
        "X10":"X11","X11":"X13","X12":"X5","X13":"X1","X14":"X2","X15":"X3"})
    shap.summary_plot(shap_values, X_train)

main()
```