

RESEARCH ARTICLE OPEN ACCESS

Predicting Environmental Violations: A Cross-Method Framework Integrating Parametric and Non-Parametric Approaches

Ashutosh Singh¹  | Salwa Saleh Almasabi² | Ajay Kumar Patel³  | Judit Petra Koltai⁴ | Honghan Qi⁵

¹Leeds University Business School, University of Leeds, Leeds, UK | ²Department of Accounting, College of Business Administration, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia | ³Institute of Management Studies (IMS), Ghaziabad, Uttar Pradesh, India | ⁴Department of Statistics, Finance and Controlling, Kautz Gyula Faculty of Business and Economics, Széchenyi István University, Győr, Hungary | ⁵University of Leicester School of Business, University of Leicester, Leicester, UK

Correspondence: Ashutosh Singh (a.singh1@leeds.ac.uk)

Received: 25 January 2026 | **Revised:** 8 July 2026 | **Accepted:** 22 July 2026

Keywords: environmental violations | firm characteristics | non-linear effects | random effects logit | SHAP values | XGBoost

ABSTRACT

Prior research has mostly relied on linear and parametric models while explaining environmental non-compliance, but they have a limited capacity to capture non-linear and asymmetric effects of elements affecting firms' environmental compliance. We integrate a random effects logit model with advanced machine learning methods on a longitudinal firm-year panel dataset of US publicly traded firms from 2000 to 2024 in order to address this gap. After comparing the random effects logit model estimates with XGBoost metrics and SHAP mean values, we find that firms' prior environmental compliance behaviour is a major predictor of future environmental penalty violations. The random effects logit model confirms the cumulative penalty count as a significant positive predictor, while XGBoost assigns it the highest gain and SHAP importance scores. The marginal contribution of the cumulative penalty count in the SHAP analyses suggests its diminishing marginal effects, which are hard to identify using the logit model alone. Cross-method comparison further reveals that governance and executive incentive variables are weak predictors of future environmental violations. The findings reveal the trajectory-dependent nature of environmental non-compliance, with enforcement outcomes strongly associated with behavioural trajectories. This study offers a methodological contribution and demonstrates the value of cross-method triangulation in resolving inconsistencies across parametric approaches by integrating parametric analysis and machine learning methods.

1 | Introduction

As large and complex datasets become more common, machine learning methods are being used more widely. However, one of the major challenges in the field of machine learning has been the development of models with high predictive accuracy (Ferreira and Gonçalves 2022). Early single decision tree methods have the limitation of overfitting by capturing noise rather than true patterns (Bramer 2020). Ensemble methods, such as gradient boosting, help improve accuracy by combining a large number of trees (Zhang and Haghani 2015). Each new tree

corrects errors that were made by earlier trees. However, traditional gradient boosting models are generally slow to train and need careful adjustment to prevent overfitting. This makes them difficult to utilise in real-world situations. The XGBoost method is developed to improve gradient boosting by penalising excessively complicated trees (P. Zhang, Jia, and Zhang 2022). This ensures that the model does not rely excessively on extremely specific patterns. The XGBoost method uses first-order and second-order gradient information to improve the efficiency of the model (Qin et al. 2021). This enables the model to converge more quickly and attain a better level of accuracy. Additionally,

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Business Strategy and the Environment* published by ERP Environment and John Wiley & Sons Ltd.

it generally takes a short amount of time for millions of observations. These characteristics make it possible for XGBoost to overcome key limitations of earlier methods. Hence, it is widely used for high-performance prediction tasks in a variety of fields.

Because of XGBoost's ability to handle large datasets, it is frequently utilised in the fields of finance and accounting for the purpose of evaluating credit risk, detecting fraud and predicting bankruptcy (Ben Jabeur et al. 2023; Hernes et al. 2023; Kumari et al. 2025). This method is widely used for supply chain optimisation and demand forecasting (Egila et al. 2025; Nsir et al. 2025). In a similar vein, XGBoost is frequently used in marketing for customer segmentation, churn prediction, pricing strategies and targeted advertising (Maqbool and Qureshi 2026). It is utilised in the field of economics for the purposes of predictive modelling, policy evaluation and macroeconomic forecasting (Nandha Gopal and Muthukaruppasamy 2024). Hence, business management has already accepted the efficiency of the XGBoost method, which supports informed strategic planning for firms.

However, the use of XGBoost in predicting the likelihood of environmental violations is still completely unexplored. Prior literature has focused on statistical methods which can provide insights into the linear relationship between the firm-related factors and environmental violations (Khan et al. 2021; Pahari et al. 2024). The traditional approaches are efficient, but they are not capable of capturing the non-linear linkages between variables. The XGBoost method is an advanced-level machine learning method that can account for the non-linear relationships within the model and has the potential to provide deeper insights into the area of environmental non-compliance. However, existing research has devoted little focus to this area, thereby constituting a substantial research gap. We focus on the following research question:

1.1 | Research Question: Which Factors Best Predict the Likelihood of Future Environmental Violations?

To address the research question, we integrate the Violation Tracker, Compustat and ExecuComp databases to create a panel dataset for US public firms from 2000 to 2024. We use a random effects logit model as a parametric baseline and calculate estimates and interpretable odds ratios associated with firm-related factors. We use the XGBoost method and SHAP values for estimating the likelihood of a violation and identifying the marginal contribution. We compare the random effects logit estimates with XGBoost importance metrics and SHAP values and obtain dominant factors responsible for future environmental violations. We find that the historical environmental penalty-related variables are major contributors to predicting future environmental violations. The gain, cover, and frequency measurements for the cumulative penalty count variable are relatively higher than those of other predictor variables. This suggests that there is a considerable path dependence in environmental non-compliance.

The XGBoost method, combined with SHAP values, can capture non-linear effects that are frequently overlooked by standard models. The study reframes the environmental infractions

conceptually as a path-dependent organisational process. This is a substantial departure from the traditional approach. Through the demonstration of declining marginal impacts of recurrent violations, the findings enhance theories of regulatory path dependence and organisational learning. The integration of different large-scale datasets into a single firm-year panel offers a comprehensive perspective on the risk of environmental penalties. The use of performance-based variables is more closely aligned with predictive accuracy and decision-making in the real world. A further methodological contribution lies in the explicit aggregation of parametric and non-parametric methods: by treating the random effects logit model and XGBoost not as alternatives but as complementary analytical tools, this study demonstrates how cross-method triangulation can resolve the inconsistencies that have long plagued the single-method parametric literature on environmental compliance. The study provides several recommendations to managers and regulators. It highlights the importance of early compliance and suggests that failure to intervene may lead to a repeated cycle. Hence, the XGBoost method can be used for predictive analytics to supplement management judgement associated with future infractions. We provide a comprehensive literature review in the next section to examine existing theoretical and empirical studies associated with environmental violations.

2 | Literature Review

2.1 | Determinants of Firms' Environmental Violations

Due to the economic, social and ecological ramifications that accompany environmental infractions committed by businesses, environmental violations continue to garner a large amount of attention in regulatory practice and policy discussions. Despite the fact that regulatory frameworks have grown over the course of the last several decades, businesses continue to violate environmental norms on a regular basis. This has prompted academics to investigate the factors that determine non-compliance. In this literature review, we outline a multi-dimensional range of factors that drive the probability of firms engaging in environmental non-compliance. These drivers include contextual, financial, governance-related, organisational and structural aspects. This literature review provides a synthesis of empirical insights that are focused on five major themes: historical path-dependence, financial constraints, operational and industry-related factors, characteristics of corporate governance, and pressures from stakeholders and regulators. These perspectives, when taken together, shed light on the reasons why environmental infractions continue to occur and which qualities of the company itself raise the likelihood of future wrongdoing.

2.2 | Historical Path-Dependence

A firm's previous non-compliance with environmental standards represents a highly reliable predictor of future regulatory offences. The prior research indicates that environmental violations follow path-dependency, and the possibility of future violations is dependent on previously committed violations (Kirk et al. 2007). Businesses frequently participate in

recurrent offences due to the fact that legal fines may not be able to sufficiently affect the routines of the organisation or the cost–benefit expectations of customers (Karpoff et al. 1999). However, previous punishments can accurately predict future infractions, despite the fact that the benefits of deterrence differ depending on the industry (Gray and Shimshack 2011). Gray and Deily (1996) conducted their study in steel plants and found that a plant's past compliance status is a strong predictor of its current compliance with air pollution regulation. Similarly, Thornton et al. (2008) demonstrate that compliance behaviour is partially formed by internal compliance cultures. These cultures are resistant to change unless there is a significant amount of pressure from the outside. When it comes to environmental responsibility, companies that have long-standing non-compliant practices internalise cost-minimisation procedures that put short-term efficiency ahead of environmental responsibility. Furthermore, they argue that companies that have previously broken environmental regulations frequently create low-trust relationships with authorities, which results in increased scrutiny but does not always eliminate wrongdoing. Managers' past learning of regulations contributes positively to path-dependency as they recognise patterns of enforcement and may formulate internal processes for which breaches can go unnoticed (Shimshack and Ward 2008). Thus, earlier violations are ongoing predictors of future actions, hinting that repeated environmental violations become ingrained when strong sanctions are absent.

2.3 | Financial Pressures and Economic Incentives

A vast amount of literature on economic theory signals that firms may reduce their focus on environmental compliance to maintain operating expenses when they encounter financial difficulties. Empirical investigations lend credence to this assertion. For example, Deily and Gray (1991) show that plants facing weaker economic conditions in a declining industry are less likely to comply with pollution regulations, and Earnhart and Lizal (2010) find that stronger economic performance does not necessarily translate into better environmental performance. Financially constrained firms may also weigh the expected penalty against the cost of compliance and choose to delay abatement investment when penalties are smaller than compliance costs (Deily and Gray 1991). At the same time, poor environmental performance is itself financially costly. Konar and Cohen (2001) show that bad environmental records reduce firms' intangible asset value, and Laplante and Lanoie (1994) document negative stock-market reactions to environmental incidents. On the other hand, companies that have access to sufficient resources may also be guilty of infractions when their plans for maximising profits prioritise cutting costs. It has been reported by Walden and Stagliano (2003) that the quality of pollution abatement information in public disclosure is low, and firms focus more on reporting financial investments. There is additional evidence in the literature that implies market rivalry influences environmental behaviour. Bansal and Roth (2000) show that firms' actions associated with ecological responsiveness are driven by legitimacy and ecological responsibility. If these compliance actions offer little competitive or legitimacy benefit, firms are less likely to

invest in them. Similarly, Delmas and Toffel (2008) demonstrate that firms' responses to environmental demands vary with the institutional pressures managers perceive. When facing weak reputational or regulatory pressure, they are more likely to under-invest in environmental management. The cost–benefit analysis of compliance is impacted according to the current financial conditions. Either because there are not enough resources to achieve the standards, or because strong financial goals drive rule-bending in order to maximise revenue, businesses may violate regulations.

2.4 | Operational and Industry Characteristics

The likelihood of a violation is strongly impacted by the complexity of operations and the structure of the industry. Higher regulatory scrutiny is implemented in pollution-intensive sectors, and even industry self-regulation initiatives in such sectors have proven ineffective without sanctions. King and Lenox (2000) show that members of the chemical industry's Responsible Care programme did not improve their environmental performance faster than non-members. Furthermore, major companies that have complicated supply chains have a difficult time monitoring the environmental performance of their subsidiaries and contractors (Hervani et al. 2005). When operations are decentralised, oversight is reduced, and the likelihood of infractions occurring, whether they are purposeful or inadvertent, is increased (Shen et al. 2025). Facilities that are embedded in big corporate networks commit more breaches due to the difficulties that arise in coordination while attempting to enforce compliance. Age and scale of the company are other important factors. Older firms tend to rely on ageing technologies and routines that resist renewal (Sørensen and Stuart 2000). Gray and Shadbegian (2005) examine the determinants of compliance for paper mills and identify that older plants comply less frequently with environmental regulations. Even small firms may not be able to comply with the environmental regulations. Tilley (2000) suggested that small firms generally lack resources and may avoid monitoring due to the fact that they are subject to less scrutiny. However, large firms have greater visibility, which makes it more likely that infractions will be discovered (Yu et al. 2017). Thus, the operational factors may act as key determinants of compliance behaviour.

2.5 | Corporate Governance and Managerial Characteristics

The organisation's governance mechanism may contribute to the establishment of environmental compliance. This mechanism is governed by the executives in the top management team, with compensation and pressure from investors. Compensation design shapes the weight managers place on environmental outcomes: short-term-oriented pay structures encourage managerial short-termism (Marinovic and Varas 2019), while Berrone and Gomez-Mejia (2009) show that in polluting industries, linking executive pay to environmental performance, particularly long-term pay, is associated with greater pollution prevention. However, some authors find that firms with more experienced executives are more likely to

implement an environmentally supportive governance structure (Dong et al. 2025). They also propose that the presence of heterogeneity in executives in the top management team is associated with improved compliance results. However, the effects of institutional ownership are not uniform. Long-term institutional investors are more likely to favour environmental compliance (Aguilera et al. 2006). On the other hand, transitory owners tend to exert pressure on companies to achieve rapid financial returns, which can sometimes result in reinforcement of environmentally harmful behaviours (Bushee 2001). Additionally, Testa et al. (2018) demonstrate that stakeholders can influence the decision to adopt environment-supporting governance mechanisms. Thus, governance factors influence the integration of environmental compliance in firms' activities, which may influence the frequency of infractions as well as the severity of those violations.

2.6 | Regulatory Enforcement, Community Pressure and Institutional Context

Regulatory enforcement is also crucial, as insufficient monitoring allows for wrongdoing to occur, whereas strict enforcement regimes minimise the number of infractions that occur (Heyes 2000). Potoski and Prakash (2005) highlight that voluntary programs (such as ISO 14001) may reduce violations when backed by robust enforcement mechanisms. Voluntary norms may merely serve as signals of compliance if they are not enforced stringently. Some studies suggest that stakeholders may pressure the firm for compliance. For example, Sharma and Henriques (2005) discovered that businesses that are confronted with requests from customers and socio-ecological stakeholders exhibit a lower tendency to violate regulations. In the emerging markets with limited regulatory capacity, Dasgupta et al. (2000) suggested that environmental management training offers a constructive complement to the limited effectiveness of traditional enforcement. The institutional environments of international organisations play an important role in environmental compliance. Vogel (2007) argues that when regulatory enforcement is relatively weak, firms may adjust their behaviour by lowering compliance efforts, resulting in compliance gaps. Additionally, Cole et al. (2006) find that multinational firms may strategically exploit disparities in environmental policies across countries. The regulatory and stakeholder contexts, therefore, are responsible for shaping compliance by defining the external consequences that are associated with ecologically hazardous activity.

Prior literature suggests a recurrent idea that environmental infractions are rarely brought about by a single element. Instead, they are a reflection of the interaction between things like the history of the organisation, the financial conditions, the complexity of the operations, the governance processes, and the pressures from the institution. When these elements are evaluated together, it is possible to provide a more precise assessment of the relative influence that they have on environmental infractions. This is in contrast to the practice of attributing non-compliance to a single determinant. Using a machine learning model, we try to identify the influence of these various aspects on environmental violations compared to a single-variable model. We follow research which advocates for integrated

frameworks to identify the extent to which each factor can affect firms' environmental violations and how firms can improve using resource allocation.

3 | Research Method

3.1 | Sample

This study makes use of a comprehensive firm-level panel dataset by combining environmental violation and penalty tracker records with Compustat financial data and ExecuComp governance data for United States publicly traded companies from 2000 to 2024. Detailed information on violations and monetary fines is given in the Violation Tracker dataset (Abebe and Acharya 2022; Shevchenko 2021). The sample is restricted to firms that appear in the Violation Tracker database, which are then matched to their corresponding Compustat and ExecuComp records. This information is aggregated to the firm-year level and connected to the accounting, financial structure, and governance features of the firms. While ExecuComp contributes executive-level information, Compustat provides other measures associated with firms' characteristics and performance (Gull et al. 2023; Singh et al. 2025; Treepongkaruna et al. 2025; Yi and Demirel 2023). The combined data is well-suited for the analysis of environmental penalty risk because it incorporates crucial managerial and performance-related metrics. The longitudinal nature of the merged datasets allows for machine learning analyses to reveal the relative importance and directional influence of these predictors.

3.2 | Variables

3.2.1 | Firm Size and Visibility

These factors influence both the regulatory exposure of a company and its ability to handle compliance requirements. Hence, they are crucial to determine environmental compliance. We measure firm size and visibility by taking the natural logarithm of total assets and sales, which reflect firm size and operational scale. We add the natural logarithm of the total number of employees to capture labour intensity.

3.2.2 | Financial Structure

The variables associated with the financial structure of firms provide a description of the capital structure, profitability, and liquidity of companies, which are indicators of their financial capability to invest in compliance-related activities and to absorb penalties. We include financial leverage measured by dividing the entire amount of long-term and short-term debt by the total assets. For the purpose of profitability, we use return on assets (ROA), which is defined as net income divided by total assets. Cash holdings are a proxy for financial flexibility and are equal to cash and short-term investments divided by total assets under consideration. Working capital intensity measures short-term liquidity in relation to the size of the company. It is calculated by dividing working capital by total assets. In order to

determine a company's short-term solvency, the current ratio is calculated by dividing current assets by current liabilities.

3.2.3 | Growth and Valuation

Certain growth features have an impact on environmental compliance, as rapidly expanding businesses may give more priority to project completion than to maintaining their environmental compliance systems. However, firms with high valuation possess strong market reputations and avoid regulatory infractions (Zou et al. 2015). We incorporate market equity, which is created by taking the log of the multiplication of the share price by the common shares outstanding. The market-to-book ratio is calculated by dividing market equity by book equity. It is a measure that represents growth prospects and intangible worth. The percentage change in sales in comparison to the previous year is referred to as sales growth, and it acts as a measurement of the dynamics of business expansion. Capital expenditure intensity is calculated by dividing total capital expenditures by total assets. This metric is used to measure investment activity.

3.2.4 | Operational Factors

Firms' operations and supply chain activities are correlated with environmental risk (Chang et al. 2021). We include firms' capital intensity as a measure of a company's dependence on fixed assets, calculated by dividing the value of the property, plant, and equipment by the total assets. Secondly, we incorporate inventory intensity, which is measured as the ratio of inventory to total assets. These factors account for operational activities, which are potentially pollution-intensive.

3.2.5 | Governance and Executive Incentives

Governance and executive features can play a crucial role in managerial decision-making and responsibility within the organisation, which may affect the regulatory compliance decision (Bajra and Čadež 2020). We include the CEO's salary in our analysis as it represents the CEO's influence on governance structures and their ability to define business strategy (Hambrick and Finkelstein 1995). The natural logarithm of total CEO compensation is used to calculate CEO pay. We include CEO tenure, measured as the number of years a CEO has held their current position, as a proxy for both experience and entrenchment. We also include the top management team (TMT) size, which is the sum of executives on the top management team. It is a reflection of the monitoring capability and governance structure of the organisation.

3.2.6 | Historical Penalty Behaviour

The past misconduct events may reflect the history and durability of enforcement. We use various proxies to capture firms' past behaviour. We create a penalty flag dummy for firms which are involved in environmental violations during a specific year. Another dummy named ever-penalised-before is created if the firm is penalised for an environmental violation prior to the

current year. We add the natural logarithms of the cumulative penalty count and sum variables, which cover the total number of yearly penalty incidents and the combined value of all penalties over time.

3.2.7 | Industry Classification

Different sectors have different levels of regulatory exposure and environmental risk profiles (Dobler et al. 2014). Hence, we use the first two-digit Standard Industrial Classification (SIC) codes for different sectors to create a sector-specific dummy, which we include as a control variable.

In order to reduce the impact of extreme outliers, all continuous variables are winsorized at the 1st and 99th percentiles (Leone et al. 2019). On the other hand, binary indicator variables are not transformed in any way. The structure of the panel is established at the firm-year level, which allows us to conduct longitudinal research on the behaviour of firms and the results of regulatory actions. Table 1 presents the definitions of variables along with the data sources.

3.3 | Estimation Method

3.3.1 | Baseline Model: Random Effects Logit

Before implementing the machine learning approach, we first estimate a random effects logit model as a baseline model. It provides coefficient estimates of each predictor, which act as a typical parametric benchmark for comparing the outcomes of the XGBoost model. The comparison between the two methods highlights the additional value of the machine learning method in terms of catching non-linear patterns that are routinely overlooked by typical parametric specifications. We consider the penalty flag as our dependent variable. We cannot prefer linear probability models for our analysis because of the binary character of the outcome (Greene 2008). The logit model estimates the log-odds of the result as a linear function of the predictors.

The longitudinal structure of the data contains unobserved firm-level heterogeneity, which is the systematic differences between companies in terms of their underlying tendency to commit environmental breaches. We have to consider this heterogeneity in the analysis in order to avoid omitted variable bias. We use a random effects logit model to address the concern of unobserved heterogeneity present in the panel dataset. The random effects measure captures the persistent unobserved heterogeneity present among organisations (Bates et al. 2024). Our objective is to calculate the conditional prediction of future violations, and the random effects estimator generates population-averaged estimates within this subgroup, which are relevant to our interest. Furthermore, our covariates consist of time-invariant and dynamic variables, which would be lost if we use the fixed effects model. Hence, the random effects logit model is more relevant for our analysis. In the analysis, all of the independent variables are measured at year t , and the results are used to make a prediction about the probability of receiving an environmental penalty in year $t + 1$. This approach is followed to address simultaneity bias

TABLE 1 | Specification of variables.

Variables	Specification	Source
<i>Firm size and visibility</i>		
Firm size	Log of total assets	Compustat
Visibility	Log of total sales	Compustat
Labour intensity	Log of number of employees	Compustat
<i>Financial structure</i>		
Leverage	Ratio of long- and short-term debts to total assets	Compustat
Profitability	Return on assets (ROA)	Compustat
Cash holdings	Ratio of cash investments to total assets	Compustat
Working capital intensity	Ratio of working capital to total assets	Compustat
Current ratio	Ratio of current assets to current liabilities	Compustat
<i>Growth and valuation</i>		
Market equity	Log of (Share price * common shares outstanding)	Compustat
Market-to-book ratio	Ratio of market equity to book equity	Compustat
Sales growth	Percentage change in sales	Compustat
Capital expenditure intensity	Ratio of total capital expenditures to total assets	Compustat
<i>Operational factors</i>		
Capital intensity	Ratio of property, plant and equipment values to total assets	Compustat
Inventory intensity	Ratio of total inventory to total assets	Compustat
<i>Governance and executive incentives</i>		
CEO salary	Log of total CEO compensation	ExecuComp
CEO tenure	Total number of years in current CEO role	ExecuComp
TMT size	Number of executives in the top management team	ExecuComp

(Continues)

TABLE 1 | (Continued)

Variables	Specification	Source
<i>Historical penalty behaviour</i>		
Penalty flag	= 1 if a firm is involved in environmental misconduct, otherwise, it is 0	Violation Tracker
Ever penalised before	= 1 if a firm is penalised for the environmental misconduct prior to the current year, otherwise, it is 0	Violation Tracker
Cumulative penalty count	Log of total number of yearly penalty incidents	Violation Tracker
Cumulative penalty sum	Log of combined value of all penalties over time	Violation Tracker

and reverse causality. This forward lag of 1 year ensures that the causal direction goes from the features of the firm to the outcomes of future regulatory actions.

3.3.2 | Machine Learning Approach

As a first step, we define the data matrix:

$$D = \{(x_i, y_i)\}_{i=1}^n \quad (1)$$

where $x_i \in R^p$ is a vector of independent variables.

$y_i \in \{0, 1\}$ is the penalty flag dummy variable.

We restrict our dataset to a specific set of variables and ensure all are numeric. The missing observations are dropped, consistent with the sample used in the baseline random effects logit model, ensuring that the XGBoost analysis is estimated on the same underlying firm-year observations as the parametric baseline so that any differences in results between the two methods reflect genuine differences in modelling approach rather than differences in the underlying sample. The model is estimated on a complete numerical feature space, and we have a total of 14,352 firm-year observations after cleaning the dataset. We separate our dependent variable from other variables and split the data chronologically into training (80%) and test (20%) samples. The training sample is used for parameter estimation, and the test sample approximates future observations. After the numerical optimisation and splitting of samples, we use the XGBoost method for the analysis. The model was configured with a maximum tree depth of 6, a learning rate of 0.05, a subsample ratio of 0.8, a column subsample ratio per tree of 0.8, a minimum child weight of 5

and a gamma value of 0.1, trained over 300 boosting rounds (Harrison 2023). A fixed random seed (1234) was set prior to both the train-test split and model training to ensure full reproducibility of the reported results. The XGBoost method uses shallow decision trees to build an additive model, and each tree enhances the previous models by targeting residual errors (Pensa et al. 2025; Simsek et al. 2025). Our dependent variable is binary, and we use a logistic model to predict log-odds, which are eventually converted into probability scores. The model equation is given by:

$$f_T(x) = \sum_{t=1}^T \eta * g_t(x) \quad (2)$$

where η is the learning rate and $g_t(x)$ is the t -th decision tree.

The logistic loss function is given by (Wang et al. 2020):

$$l(y_i, \hat{p}_i) = - [y_i * \log(\hat{p}_i) + (1 - y_i) * \log(1 - \hat{p}_i)] \quad (3)$$

The XGBoost method tries to minimise this loss function. The model is used on the training sample and predicted probability scores are assigned to the test sample. Although the model provides original probabilities, we use a cutoff rule to map probabilities into class labels (Harrison 2023).

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{p}_i \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

We use Shapley additive explanations (SHAP) to assign each independent variable a contribution value and show its predictive power with respect to the baseline prediction (Mangalathu et al. 2020). The SHAP decomposition function is given by:

$$f(x_i) = \theta_0 + \sum_{j=1}^p \theta_j(x_i) \quad (5)$$

where θ_0 is the baseline expected value and $\theta_j(x_i)$ is the variable j 's contribution.

4 | Results

After removing all the missing values from the merged dataset, we have a total of 14,352 observations from 2000 to 2024. Table 2 provides the descriptive statistics of all the variables used in the analysis. There is substantial variation across all of the variables in the dataset, and some variables have a higher standard deviation in comparison to their respective mean values. Table 2 indicates that mean values capture central tendencies, while several variables exhibit substantial variability.

Figure 1 shows the correlation matrix for the variables. The majority of pairwise correlations are of a modest magnitude, which indicates that multicollinearity is not a big concern for the majority of the variables. However, some variables have strong positive correlations. The high values of penalty-related measures reflect continuity in regulatory enforcement over time. On the other hand, the majority of other predictors have relatively poor correlations. Based on these patterns, the predictors are able to capture distinct parts of the characteristics of the firm, while also identifying clusters of closely related variables.

TABLE 2 | Descriptive statistics.

Variables	Mean	SD	Sample
<i>Firm size and visibility</i>			
Firm size	8.32	1.80	14,352
Visibility	8.01	1.62	14,352
Labour intensity	2.19	1.45	14,352
<i>Financial structure</i>			
Leverage	0.31	0.20	14,352
Profitability	0.04	0.09	14,352
Cash holdings	0.09	0.10	14,352
Working capital intensity	0.15	0.16	14,352
Current ratio	1.91	1.27	14,352
<i>Growth and valuation</i>			
Market equity	7.26	2.99	14,352
Market-to-book ratio	3.39	4.31	14,352
Sales growth	0.09	0.25	14,352
Capital expenditure intensity	0.06	0.05	14,352
<i>Operational factors</i>			
Capital intensity	0.65	0.41	14,352
Inventory intensity	0.12	0.12	14,352
<i>Governance and executive incentives</i>			
CEO salary	8.54	0.99	14,352
CEO tenure	7.60	6.76	14,352
TMT size	5.72	1.12	14,352
<i>Historical penalty behaviour</i>			
Penalty flag	0.28	0.45	14,352
Ever-penalised-before	0.68	0.47	14,352
Cumulative penalty count	0.95	0.86	14,352
Cumulative penalty sum	8.14	5.90	14,352

Table 3 shows the results of the analysis of the random effects logit model. The second column shows log odds, and we convert them to odds ratios, which are presented in the third column of Table 3. The odds ratio provides a meaningful representation and signals the importance of predictors. The firm size variable under the firm size and visibility category is positive and significant ($\beta = 0.27, p < 0.05$). This indicates that a one-unit increase in log total assets is associated with 31% higher odds of receiving an environmental penalty. The coefficient of visibility is positive and significant ($\beta = 0.37, p < 0.01$). This suggests that a one-unit increase in the log of total sales is associated with 45% higher odds of obtaining an environmental penalty. The labour intensity estimates are not significant ($\beta = -0.06, p > 0.1$). Regarding factors related to the financial structure, we find that only the cash holdings variable is negative and significant ($\beta = -1.82, p < 0.01$). Hence, a greater cash liquidity has 84% lower odds of receiving an environmental

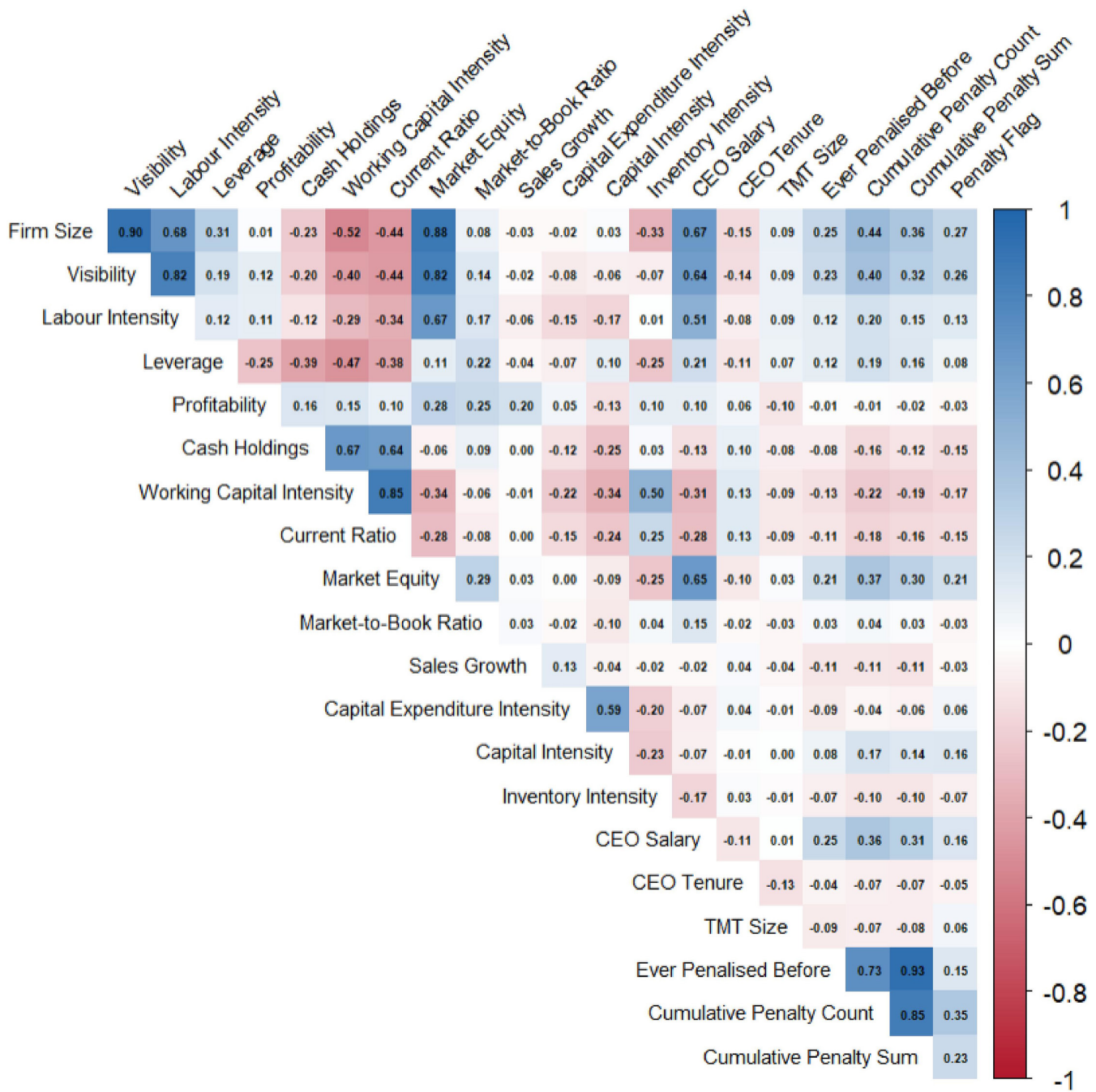


FIGURE 1 | Correlation diagram.

penalty. Under the growth and valuation category, only capital expenditure intensity is positive and significant ($\beta = 1.90$, $p < 0.05$). The odds ratio of 6.70 suggests that a higher capital investment intensity is at a greater risk of incurring environmental penalties as compared to firms with lower capital investment intensity. The capital intensity variable is positive and significant under operational factors ($\beta = 0.43$, $p < 0.01$). Firms with more reliance on capital assets have 54% higher odds of receiving penalties. We do not find any significance in the variables present in the governance and executive incentives category. In the historical penalty behaviour category, the ever penalised before dummy is negative and significant ($\beta = -0.35$, $p < 0.1$). The odds ratio of 0.70 suggests that firms with a history of environmental penalty have 30% lower odds

of receiving a penalty in the current year, which signals that these firms perform corrective action or internal compliance reform to avoid repeating the same violations. The cumulative penalty count is positive and significant ($\beta = 0.27$, $p < 0.05$). The odds ratio of 1.30 indicates that there will be a 30% increase in the odds of receiving a future penalty for an additional unit of environmental penalty. The cumulative penalty sum variable is not significant ($\beta = -0.003$, $p > 0.1$). We use robust standard errors to address the heteroskedasticity issue. The VIF value is less than 10, indicating that the analysis is free from multicollinearity.

The Receiver Operating Characteristic (ROC) curve for the XGBoost model is illustrated in Figure 2. The ROC curve shows

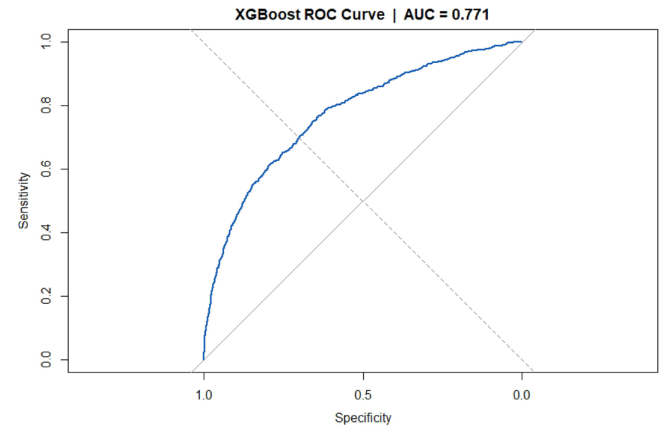
TABLE 3 | Random effects panel logit analysis.

Variables	Estimates (SE)	Odds ratios
<i>Firm size and visibility</i>		
Firm size	0.27** (0.13)	1.31**
Visibility	0.37*** (0.12)	1.45***
Labour intensity	−0.06 (0.08)	0.95
<i>Financial structure</i>		
Leverage	−0.09 (0.30)	0.91
Profitability	−0.47 (0.42)	0.62
Cash holdings	−1.82*** (0.55)	0.16***
Working capital intensity	0.44 (0.60)	1.55
Current ratio	0.01 (0.05)	1.01
<i>Growth and valuation</i>		
Market equity	−0.03 (0.06)	0.97
Market-to-book ratio	−0.01 (0.01)	0.99
Sales growth	−0.15 (0.12)	0.86
Capital expenditure intensity	1.90** (0.90)	6.70**
<i>Operational factors</i>		
Capital intensity	0.43*** (0.15)	1.54***
Inventory intensity	−0.82 (0.73)	0.44
<i>Governance and executive incentives</i>		
CEO salary	−0.01 (0.04)	0.99
CEO tenure	0.00 (0.04)	1.00
TMT size	0.01 (0.02)	1.01
<i>Historical penalty behaviour</i>		
Ever-penalised-before	−0.35* (0.21)	0.70*

(Continues)

TABLE 3 | (Continued)

Variables	Estimates (SE)	Odds ratios
Cumulative penalty count	0.27** (0.12)	1.30**
Cumulative penalty sum	−0.003 (0.02)	0.99

*** $p < 0.01$.** $p < 0.05$.* $p < 0.10$.**FIGURE 2** | ROC curve of the XGBoost classifier.

the relationship between sensitivity, which is the proportion of accurately recognised future violators, and specificity, which is the proportion of non-violators correctly identified as non-violators, across all of the available probability thresholds. The value of the area under the curve (AUC) is 0.771, which indicates that it accurately differentiates between future violators and non-violators around 77% of the time on unseen data. Hence, the XGBoost model is able to successfully identify regular patterns in the relationship between the characteristics of the firm and the potential likelihood of environmental penalties in the future.

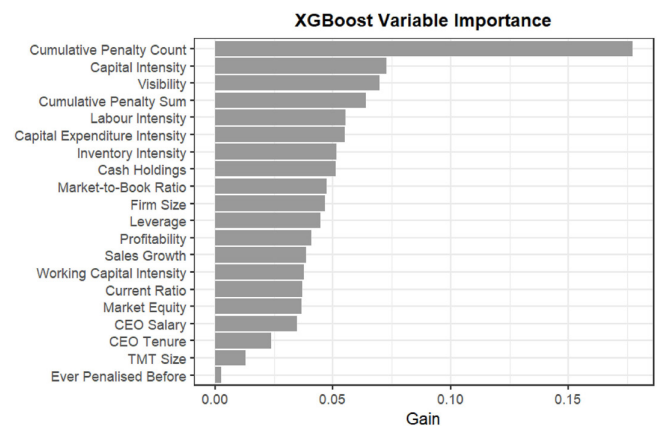
Table 4 shows the relative importance of each predictor variable in the XGBoost model using gain, cover and frequency metrics (Chen et al. 2019; Y. Zhang, Shi, et al. 2022). The accumulated history of environmental penalties is crucial for forecasting future violations, as suggested by the three metrics for the cumulative penalty count variable (gain = 0.18, cover = 0.07, frequency = 0.03). In addition, the cumulative penalty sum makes a meaningful contribution, with gain, cover and frequency values of 0.06, 0.05 and 0.05. However, the three metrics are lowest for the ever-penalised-before dummy variable (gain = 0.003, cover = 0.002, frequency = 0.004), suggesting that a binary variable for prior enforcement carries far less predictive information than the other two historical penalty variables. With regard to the remaining categories of variables, operational factors and more specifically capital intensity (gain = 0.07, cover = 0.07, frequency = 0.07) demonstrate a relatively stronger contribution in comparison to financial structure, growth and valuation, and firm size, while performing on par with visibility. Governance

TABLE 4 | XGBoost variable importance measures.

Variables	Gain	Cover	Frequency
<i>Firm size and visibility</i>			
Firm size	0.05	0.05	0.04
Visibility	0.07	0.08	0.06
Labour intensity	0.06	0.06	0.07
<i>Financial structure</i>			
Leverage	0.04	0.06	0.06
Profitability	0.04	0.06	0.06
Cash holdings	0.05	0.06	0.06
Working capital intensity	0.04	0.04	0.05
Current ratio	0.04	0.06	0.05
<i>Growth and valuation</i>			
Market equity	0.04	0.04	0.05
Market-to-book ratio	0.05	0.05	0.06
Sales growth	0.04	0.05	0.06
Capital expenditure intensity	0.06	0.06	0.07
<i>Operational factors</i>			
Capital intensity	0.07	0.07	0.07
Inventory intensity	0.05	0.06	0.07
<i>Governance and executive incentives</i>			
CEO salary	0.03	0.04	0.05
CEO tenure	0.02	0.03	0.03
TMT size	0.01	0.02	0.01
<i>Historical penalty behaviour</i>			
Ever-penalised-before	0.003	0.002	0.004
Cumulative penalty count	0.18	0.07	0.03
Cumulative penalty sum	0.06	0.05	0.05

and executive incentive factors continuously record the lowest importance values among all non-penalty predictors, with TMT size in particular reporting gains of only 0.01 and a frequency of 0.01. Hence, Table 4 demonstrates that past penalty behaviour and operational variables are more important factors than financial, valuation and governance aspects in the XGBoost model.

Comparing the values of the XGBoost metrics with the random effects logit estimates reveals both convergence and considerable divergence between the two approaches. This demonstrates that the machine learning framework brings additional explanatory value to the analysis. Both models signal the importance of the cumulative penalty count in predicting future environmental violations. Similarly, they also show that company size, visibility and capital intensity are positively linked with penalty risk. However, the relative importance of these factors is more modest in the XGBoost framework than the logit coefficients alone might

**FIGURE 3** | Graphical representation of XGBoost variable importance results.

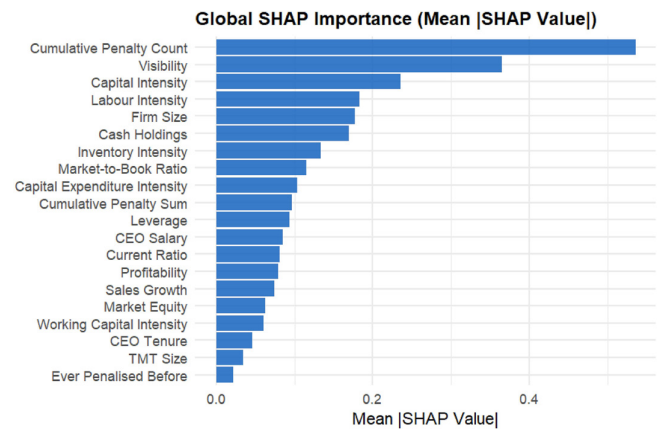
suggest. Variables like cash holdings and capital expenditure intensity are significant in the logit model but rank only in the middle tier of XGBoost importance scores. The parametric model may partly reflect their genuine superiority. On the other hand, the XGBoost model shows more importance for operational factors like inventory intensity. However, they are not significant in the logit model, suggesting that these variables are not detected in the parametric specification. Convergent data suggest that governance structures have a weak association with enforcement outcomes. This is supported by the fact that governance and executive variables continue to be weak across both methodologies, with no significant logit coefficients and the lowest XGBoost significance scores. The ever-penalised-before dummy is significant at the 10% level in the logit model, but it shows less importance in the XGBoost model. These results suggest that the observed effect may be a model-specific artefact of the random effects logit model. Thus, the XGBoost framework uncovers a more nuanced structure of predictors and shows some alignment with the estimates of the random effects logit model. Figure 3 depicts the graphical representation of these results.

The results of the SHAP analysis are reported in Table 5. The mean values are based on game-theoretic Shapley values, which allow for measuring the average marginal contribution of each variable to individual forecasts (Henna et al. 2025). In contrast to gain, cover and frequency metrics of the XGBoost model, SHAP values reflect the variable's actual influence on the model's output at the observation level. Hence, SHAP values can provide a more theoretically meaningful picture of the importance of variables. After comparing mean SHAP values with the random effects logit model's estimates, we discover that both analyses identify the cumulative penalty count as the most powerful predictor of future environmental violations. The mean values of visibility and firm size variables are ranked second and fourth in Table 5 (0.36 and 0.18). These variables are also significant in the logit analysis. Capital intensity, which ranks third (0.24), continues to be highly significant in the random effects logit model. However, labour and inventory intensities are ranked fourth (0.18) and seventh (0.13) in Table 5, but they are not significant in the logit analysis. These variables may carry significant non-linear contributions that the parametric specification is unable to identify. Capital expenditure intensity is significant in the logit framework but ranked ninth in the SHAP framework

TABLE 5 | Mean absolute SHAP values.

Variables	Mean value
<i>Firm size and visibility</i>	
Firm size	0.18
Visibility	0.36
Labour intensity	0.18
<i>Financial structure</i>	
Leverage	0.09
Profitability	0.08
Cash holdings	0.17
Working capital intensity	0.06
Current ratio	0.08
<i>Growth and valuation</i>	
Market equity	0.06
Market-to-book ratio	0.11
Sales growth	0.07
Capital expenditure intensity	0.10
<i>Operational factors</i>	
Capital intensity	0.24
Inventory intensity	0.13
<i>Governance and executive incentives</i>	
CEO salary	0.09
CEO tenure	0.05
TMT size	0.03
<i>Historical penalty behaviour</i>	
Ever-penalised-before	0.02
Cumulative penalty count	0.54
Cumulative penalty sum	0.10

(0.10), suggesting that the logit may exaggerate its significance. The cash holdings variable is significant in the random effects logit model and ranks sixth in SHAP analysis (0.17). Both analyses suggest that cash holdings are strongly associated with environmental penalties. The ever-penalised-before binary variable is weakly significant at the 10% level, and the SHAP mean value (0.02) also signals the limited contribution of this variable. Governance and executive incentive variables are not significant in the logit analysis, and their values are also among the lowest in the SHAP analysis (0.09, 0.05 and 0.03). Hence, these variables have a negligible influence on the outcomes of enforcement across both methodologies. The XGBoost model achieves an AUC of 77%, but it does not fully explain how and in what direction individual variables drive forecasts for specific organisations. SHAP analysis is a further extension of the XGBoost model, which is able to transform predictive accuracy into interpretable insights. SHAP values offer a more comprehensive

**FIGURE 4** | Graphical representation of mean SHAP values.

picture of variable importance and help to capture non-linear and asymmetric effects that may not be identified by the logit model. Figure 4 provides a pictorial representation of the mean absolute SHAP values.

Figure 5 shows the observation-level visualisation of each variable's contribution to the predicted penalty risk. Each observation is represented in the horizontal position, which indicates both the direction and the magnitude of the variable's SHAP value. The positive values (right of zero) increase the predicted probability of an environmental penalty, while negative values (left of zero) decrease it. The cumulative penalty count has the highest value, followed by visibility, capital intensity, labour intensity and firm size. Governance variables such as TMT size, CEO tenure and ever-penalised-before dummy appear at the bottom of the figure, confirming their negligible marginal contributions. The mean absolute SHAP value is reported alongside each variable on the left side of the figure. There is a colour scale on the right side of the figure that represents the feature value of each variable for a specific observation. The colour purple indicates a high feature value, the colour green represents a mid-range value and the colour yellow indicates a low feature value. This allows us to observe both the direction of the SHAP contribution and whether it is driven by high or low values of that variable. The cumulative penalty count variable has purple dots extending far to the right, which signals that firms with a high cumulative penalty history receive strongly positive SHAP values. On the other hand, yellow dots clustered to the left of zero indicate that firms with little enforcement history carry a negative contribution from this variable.

We provide the marginal effect of the cumulative penalty count in Figure 6. This plot highlights the non-linear and saturating relationship that exists between the cumulative penalty count and the likelihood of environmental penalties in the future. This figure separately shows the trend of the marginal contribution of the cumulative penalty count to penalty risk. The horizontal axis displays the cumulative penalty count variable, and the vertical axis displays the variable's contribution to the predicted penalty risk for that particular observation. Each vertical line in the figure serves as a representation of a single firm-year observation. Using the colour scale, the same information is reinforced, with blue vertical lines representing companies that have low cumulative penalty counts and red

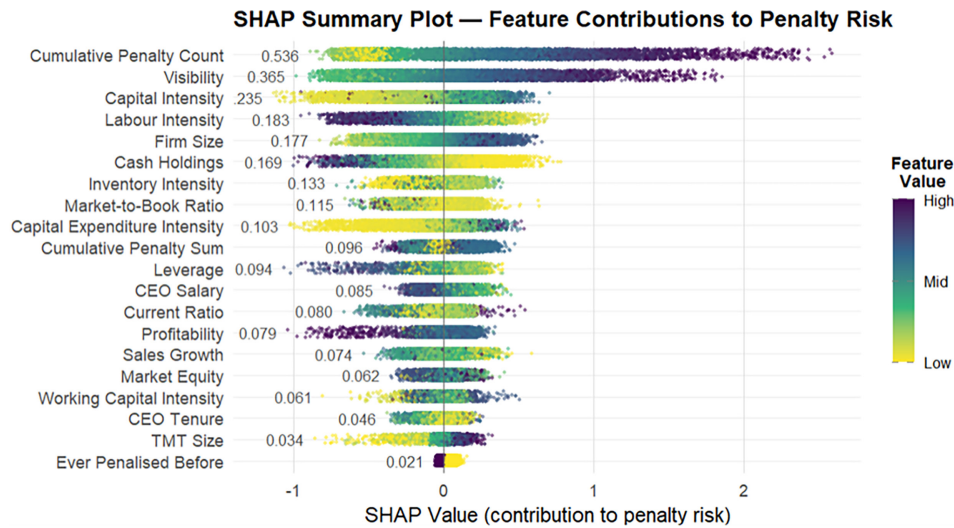


FIGURE 5 | Observation-level SHAP contributions.

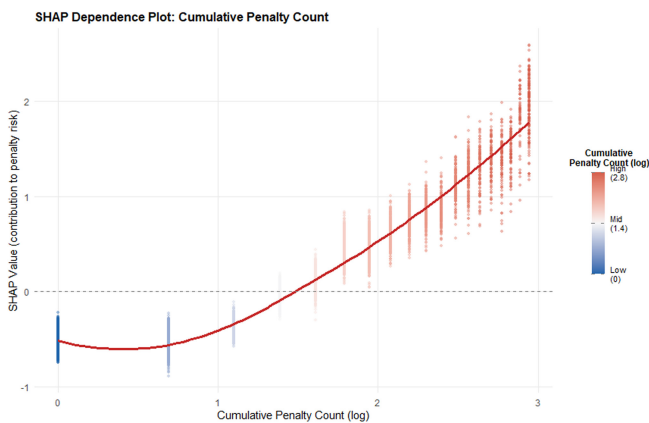


FIGURE 6 | Marginal effect of the cumulative penalty count on model predictions.

vertical lines representing companies that have large cumulative counts. The smooth red trend line reveals three distinct patterns. First, firms with little penalty history have negative SHAP values, and their enforcement history actually reduces their predicted penalty risk. Second, there exists a threshold value of the cumulative penalty count for which SHAP values shift from negative to positive. Third, we can observe the diminishing marginal effect of the cumulative penalty count on the right side after extending the trend, where the rate of increase in SHAP values begins to flatten at very high levels of cumulative penalties. These outcomes could not be produced by a standard linear or logit model.

5 | Discussion and Conclusion

Our research focuses on the driving force behind the likelihood of future environmental violations. Our investigation addresses one of the most enduring puzzles in the field of environmental governance studies at this point. The identification of the organisational, financial and governance-related traits that consistently signal potential non-compliance has been a focus of investigation for academics and regulatory authorities for several decades.

Nevertheless, in spite of the fact that there is a considerable body of research on parametric models, managers, investors and enforcement agencies in the real world continue to find these models to have minimal practical utility. We demonstrate that the existing parametric approaches may be incomplete due to their structural limitations. This incompleteness is not a reflection of the true complexity of environmental non-compliance, as the present study demonstrates. We combine parametric and non-parametric methodological traditions in order to arrive at a more comprehensive analysis of the elements that are responsible for environmental violations. The random effects logit model and the XGBoost machine learning method are not competing tools in this context. They are complementary lenses that have been trained on the same data. The logit model coefficients' significance and interpretations are based on the premises of linearity and additive separability, which the real world does not honour. The parametric models may not reflect the feedback mechanisms and path-dependent trajectories, which are present in the dynamics of regulatory non-compliance. Thus, parametric models may provide incomplete conclusions. The XGBoost and SHAP analyses allow the researchers to confirm the findings of the parametric baseline model.

We discover that the variables associated with the historical penalty behaviour show the highest dominance, and the finding is consistent across both approaches. It has been determined by both the random effects logit model and the XGBoost framework that the cumulative penalty count is the most accurate indicator of future environmental breaches. The logit model suggests that a unit increase in the cumulative penalty count is linked to a 30% increase in the odds of obtaining a future violation. The gain metric for the cumulative penalty count in the XGBoost model is much higher than that of all other predictors. Additionally, the mean SHAP value of the cumulative penalty count has the highest rank compared to other variables. The integration of parametric and non-parametric methods is a robust approach for the path-dependent nature of environmental non-compliance. Our findings extend the earlier work of Kirk et al. (2007) and Thornton et al. (2008), which established that compliance behaviour is shaped by internal organisational cultures that resist change in the absence of substantial external pressure. The present study makes an important contribution by revealing,

through the use of SHAP dependence plots, that this relationship is non-linear and saturating. The findings of this study suggest that companies with limited enforcement histories have negative SHAP contributions. Their limited history is associated with a reduced predicted risk of being penalised. There is a threshold at which the marginal SHAP value turns positive, and at very high cumulative counts, the rate of increase flattens out significantly. This pattern of diminishing marginal effect is completely undetectable using the logit and linear regression frameworks. This effect is detected with the incorporation of machine learning. It suggests that the enforcement agencies could prioritise firms approaching the threshold region, where the marginal risk of future violations increases most sharply.

We obtain a more nuanced picture of the variables included in financial structure, growth and valuation categories. The logit model reveals that cash holdings are a significant negative predictor. These findings align with the prior literature (Konar and Cohen 2001; Laplante and Lanoie 1994), which states that financial constraints restrict compliance investment. Additionally, the logit model shows that the capital expenditure intensity may be associated with a higher level of environmental exposure. However, it is necessary to exercise caution when interpreting these data in light of the results obtained by XGBoost. In terms of mean absolute SHAP values, cash holdings and capital expenditure intensity obtain relatively lower ranks. Both of these rankings are lower than those of the cumulative penalty count and operational factors. The overstatement of the marginal contribution of financial structure variables shown by the random effects logit is mismatched with the XGBoost model. When the genuine predictive weight of historical punishment behaviour is taken into account and when it is permitted to interact non-linearly with other variables, the relative value of financial structure variables decreases. These variables contribute relatively little predictive information about the regulatory trajectory of a company. Hence, the economic models of compliance that consider financial structure variables as major explanatory elements need more scrutiny while focusing on environmental violations.

The logit model demonstrates that firm size and visibility are both significant and positive predictors, which aligns with prior research (Yu et al. 2017). Firm size and visibility also have a high ranking in the SHAP framework. However, the variable labour intensity is not significant in the logit model but is ranked fourth in the SHAP analysis. One possible explanation is that more labour-intensive companies may be more observable to regulators, which could enhance the probability of enforcement regardless of actual compliance conduct. While the logit model tries to capture this, it does not take into account the non-linear interactions that the XGBoost model encodes between firm size, operational factors and historical penalty behaviour. The capital intensity variable shows its dominance in both methodologies. It suggests that operational characteristics are responsible for the creation of structural exposure to environmental risk. The variables associated with governance and executive incentives do not signal their importance in random effects logit and XGBoost models. As a result of this alignment across both methodologies, there is substantial evidence that governance structures have a negligible direct influence on enforcement outcomes. However, prior literature establishes a connection between environmental performance and executive incentives (Aguilera et al. 2006;

Dong et al. 2025). This suggests that boards may adopt environmental policies and signal their commitment to the environment, but the actual probability of regulatory violation appears to be driven much more powerfully by the accumulated behavioural trajectory of the company. The governance mechanisms may indirectly influence the path-dependent enforcement dynamics through broad strategic orientations. Furthermore, the ever-penalised-before binary variable is significant in the logit model but not a dominant factor in the XGBoost model. Thus, a simple binary indicator of past violation carries less information, and the nuance of historical penalty behaviour is more important.

Hence, we address the real-world concern regarding the inconsistency of the parametric environmental compliance research, which may bring inconsistent and realistically limited outcomes. The parametric tradition is not incorrect. It is lacking in several respects. The problem is that it assumes linearity when there is none. It gives significance to variables that rank poorly in non-parametric importance frameworks and are unable to recover the path-dependent dynamics. We do not recommend treating parametric and non-parametric approaches as alternatives. Our study illustrates that combining parametric and non-parametric approaches generates a more comprehensive and practically actionable understanding of environmental non-compliance.

5.1 | Contributions

The majority of researchers rely on linear regression models while working on environmental violations. We contribute to the prior literature by adopting an advanced machine learning approach in this field of academic research. The traditional regression methods have several assumptions which may be hard to follow during the analysis and interpretation. However, it is unlikely to assume that the regulatory environment is characterised by a linear relationship between dependent and independent variables. Hence, it is crucial to use a machine learning approach, which accounts for non-linear relationships. Furthermore, this is more relevant in the context of environmental violations, where business characteristics may be contingent on previous enforcement histories. This demonstrates the methodological limits of conventional techniques in this research domain, which are highlighted by the findings. A critical methodological innovation of this study is the deliberate aggregation of a parametric benchmark with a non-parametric benchmark. The XGBoost method is capable of capturing the true contributions of predictors. In addition, the SHAP values reveal the patterns of diminishing marginal effects of predictors. Hence, it offers a more nuanced empirical operationalisation of path dependence. The comparison allows the identification of the overestimation of the logit coefficients of certain variables in the parametric models. This is a methodologically distinctive triangulation, which offers a principled solution to the longstanding puzzle of why single-method parametric studies have produced inconsistent findings in this literature.

We provide a contribution to the field of methodology with the conceptualisation and evaluation of variable significance values. When conducting traditional regression-based research, it is common practice to infer significance based on statistical significance or the magnitude of the coefficient. Both of these

factors are sensitive to the model specification and the sample size. However, we use gain, cover and frequency metrics, which are associated with performance-based evaluation and relevant to the outcomes of complex organisational situations. The comparison between these performance-based XGBoost metrics and the odds ratios generated by the random effects logit model reveals important divergences that have substantive theoretical implications. For instance, the ever penalised before binary dummy is weakly significant in the logit model but registers near-zero importance in all XGBoost metrics and SHAP analysis. Thus, the logit results may be a model-specific artefact rather than a genuine predictor of future violations. This kind of cross-method diagnostic is only possible when parametric and non-parametric results are explicitly compared, a practice this study establishes as a standard for future research in this domain. Our analysis encourages a data-informed assessment, which is becoming increasingly essential in environmental governance disciplines. Environmental compliance can be characterised by historical compliance behaviours along with other firm-related factors. Our approach captures the asymmetric effects of predictors by showing the diminishing marginal effects of predictors. Additionally, we recommend relying less on static factors for researching regulatory phenomena, which are uneven by nature. We contribute to the pluralism of methodologies within this field by illustrating how computational methods can be incorporated into organisational research. We position machine learning as a supplementary instrument to improve empirical precision.

From a managerial perspective, firms should consider environmental violations as persistent organisational trajectories instead of discrete correctable incidents. They should keep in mind that early-stage compliance breaches may bring the risk of persistent enforcement. Hence, it is necessary to comply during the early phases of organisational growth and expansion into activities that are environmentally sensitive. Short-term remedies may not be sufficient to change the regulatory trajectory of the company. Imposing an environmental violation penalty may also bring brand damage, which is not beneficial to firms. Thus, managers should think about environmental compliance as a long-term strategic issue and incorporate efficient monitoring systems and internal accountability structures within the organisation. Addressing the core issues may shift the attention from discovery-based enforcement to confirmation-based enforcement. This trajectory-based view of violations is supported by both the random effects logit model and the XGBoost framework. Managers should consider their firm's regulatory history as the single most informative signal of future enforcement exposure. They should be aware that regulatory attention is dynamic in nature, and the current improved policies may not be helpful in the presence of a substantial violation history. They should maintain consistency in environmental performance as well as the hazards that are associated with varying their efforts to comply with regulations. Environmental compliance risk should be seen as an essential component of firm risk management rather than a peripheral concern. Managers should integrate environmental compliance indicators while assessing risks associated with environmental violations. This integration has the potential to enhance internal awareness and enable better-informed strategic planning. Though

financial resources play a secondary role, the data suggest that they may continue to influence compliance outcomes. Managers should realise that financial resources alone may not improve environmental performance if firms have a previous history of environmental violations. This discrepancy in the dominance of cash holdings gives a signal to managers that they should not focus on this variable for resetting the regulatory trajectory.

5.2 | Limitations and Future Research Directions

Though we provide crucial insights, our study has some limitations which draw attention to areas of research that have the potential to provide further discoveries. When these limitations are addressed, it is possible to further enhance theoretical understanding and methodological techniques within the literature of environmental governance and organisational misbehaviour. We investigate solely publicly traded US firms by merging different datasets. Although this integrated dataset offers a wealth of information at the firm level and makes it possible to conduct longitudinal research, it restricts the extent to which the findings may be generalised. Researchers can collect the dataset from private firms and perform a combined dataset analysis to get additional insights. The scope can be expanded to developed or emerging markets, which have different regulatory enforcement. The path-dependent dynamics should be investigated through comparative studies, as observed in this study.

We measure environmental violations with the help of reported monetary penalties. Researchers can collect data on regulatory warnings and perform a study to understand the nuances of regulatory scrutiny. Due to the high predictive capacity of historical penalties, companies having a history of infractions may be subjected to a higher level of inspection. Hence, subsequent research can be performed to identify the impacts of inspection frequency to untangle behavioural persistence from monitoring effects. A quasi-experimental technique of causal inference may be of use in elucidating the directionality of these interactions. We use year-wise aggregation of infractions and company characteristics. Short-term dynamics (monthly or quarterly) may also provide additional insights into our study. There is potential for future research to make use of data with a higher frequency, such as recordings at the quarterly or event level, in order to investigate how businesses modify their behaviour promptly following enforcement actions. When it comes to recording short-term organisational reactions and differentiating between symbolic and real changes in environmental practices, event study methods may prove to be particularly beneficial. Finally, future research can be conducted on specific industries, and researchers can provide additional insights with respect to each industry.

Author Contributions

Ashutosh Singh: conceptualisation, investigation, writing – original draft, methodology, writing – review and editing, software, formal analysis, project administration. **Salwa Saleh Almasabi:** conceptualisation, supervision, writing – original draft, validation, project administration. **Ajay Kumar Patel:** conceptualisation, writing – original draft,

validation, writing – review and editing, project administration. **Judit Petra Koltai:** supervision, writing – review and editing, validation, conceptualisation. **Honghan Qi:** data curation, writing – review and editing, validation, investigation.

Funding

This work is supported by the Princess Nourah bint Abdulrahman University Researcher Supporting Project (Project No. PNURSP2026R860), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

- Abebe, M. A., and K. Acharya. 2022. "Founder CEOs and Corporate Environmental Violations: Evidence From S&P 1500 Firms." *Business Strategy and the Environment* 31, no. 3: 1204–1219.
- Aguilera, R. V., C. A. Williams, J. M. Conley, and D. E. Rupp. 2006. "Corporate Governance and Social Responsibility: A Comparative Analysis of the UK and the US." *Corporate Governance: An International Review* 14, no. 3: 147–158.
- Bajra, U., and S. Čadež. 2020. "Alternative Regulatory Policies, Compliance and Corporate Governance Quality." *Baltic Journal of Management* 15, no. 1: 42–60.
- Bansal, P., and K. Roth. 2000. "Why Companies Go Green: A Model of Ecological Responsiveness." *Academy of Management Journal* 43, no. 4: 717–736.
- Bates, M. D., L. E. Papke, and J. M. Wooldridge. 2024. "Non Linear Correlated Random Effects Models With Endogeneity and Unbalanced Panels." *Econometric Reviews* 43, no. 9: 713–732.
- Ben Jabeur, S., N. Stef, and P. Carmona. 2023. "Bankruptcy Prediction Using the XGBoost Algorithm and Variable Importance Feature Engineering." *Computational Economics* 61, no. 2: 715–741.
- Berrone, P., and L. R. Gomez-Mejia. 2009. "Environmental Performance and Executive Compensation: An Integrated Agency-Institutional Perspective." *Academy of Management Journal* 52, no. 1: 103–126.
- Bramer, M. 2020. "Avoiding Overfitting of Decision Trees." In *Principles of Data Mining*, 121–136. Springer London.
- Bushee, B. J. 2001. "Do Institutional Investors Prefer Near-Term Earnings Over Long-Run Value?" *Contemporary Accounting Research* 18, no. 2: 207–246.
- Chang, X., Y. Huang, M. Li, X. Bo, and S. Kumar. 2021. "Efficient Detection of Environmental Violators: A Big Data Approach." *Production and Operations Management* 30, no. 5: 1246–1270.
- Chen, T., T. He, M. Benesty, and V. Khotilovich. 2019. "Package 'XGBoost'." *R Version* 90 1–66: 1.
- Cole, M. A., R. J. Elliott, and P. G. Fredriksson. 2006. "Endogenous Pollution Havens: Does FDI Influence Environmental Regulations?" *Scandinavian Journal of Economics* 108, no. 1: 157–178.
- Dasgupta, S., H. Hettige, and D. Wheeler. 2000. "What Improves Environmental Compliance? Evidence From Mexican Industry." *Journal of Environmental Economics and Management* 39, no. 1: 39–66.
- Deily, M. E., and W. B. Gray. 1991. "Enforcement of Pollution Regulations in a Declining Industry." *Journal of Environmental Economics and Management* 21, no. 3: 260–274.
- Delmas, M. A., and M. W. Toffel. 2008. "Organizational Responses to Environmental Demands: Opening the Black Box." *Strategic Management Journal* 29, no. 10: 1027–1055.
- Dobler, M., K. Lajili, and D. Zéghal. 2014. "Environmental Performance, Environmental Risk and Risk Management." *Business Strategy and the Environment* 23, no. 1: 1–17.
- Dong, H., C. Zhang, and W. Teng. 2025. "Top Management Team Heterogeneity and Corporate ESG Performance." *Finance Research Letters* 73: 106610.
- Earnhart, D., and L. Lizal. 2010. "Effect of Corporate Economic Performance on Firm-Level Environmental Performance in a Transition Economy." *Environmental and Resource Economics* 46, no. 3: 303–329.
- Egila, A. E., M. M. Kamal, S. Kumar Mangla, N. Rich, and B. Tjahjono. 2025. "Highly Reliable Organisations and Sustainability Risk Management: Safety Cultures in the Nigerian Oil and Gas Supply Chain Sector." *Business Strategy and the Environment* 34, no. 2: 2680–2701.
- Ferreira, C., and G. Gonçalves. 2022. "Remaining Useful Life Prediction and Challenges: A Literature Review on the Use of Machine Learning Methods." *Journal of Manufacturing Systems* 63, no. May: 550–562.
- Gray, W. B., and M. E. Deily. 1996. "Compliance and Enforcement: Air Pollution Regulation in the US Steel Industry." *Journal of Environmental Economics and Management* 31, no. 1: 96–111.
- Gray, W. B., and R. J. Shadbegian. 2005. "When and Why Do Plants Comply? Paper Mills in the 1980s." *Law & Policy* 27, no. 2: 238–261.
- Gray, W. B., and J. P. Shimshack. 2011. "The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence." *Review of Environmental Economics and Policy* 5, no. 1: 3–24. <https://doi.org/10.1093/reep/req017>.
- Greene, W. H. 2008. "The Econometric Approach to Efficiency Analysis." *Measurement of Productive Efficiency and Productivity Growth* 1, no. 1: 92–250.
- Gull, A. A., N. Hussain, S. A. Khan, R. Mushtaq, and R. Orij. 2023. "The Power of the CEO and Environmental Decoupling." *Business Strategy and the Environment* 32, no. 6: 3951–3964.
- Hambrick, D. C., and S. Finkelstein. 1995. "The Effects of Ownership Structure on Conditions at the Top: The Case of CEO Pay Raises." *Strategic Management Journal* 16, no. 3: 175–193.
- Harrison, M. 2023. *Effective XGBoost: Tuning, Understanding, and Deploying Classification Models*. Matt Harrison.
- Henna, S., M. R. Sakhamuri, L. G. Moitra, and U. Rathnayake. 2025. "Game-Theoretic Explainable AI for Ensemble-Boosting Models in Early Malware Prediction for Computer Systems." *International Journal of Computational Intelligence Systems* 18, no. 1: 318.
- Hernes, M., J. Adaszyński, and P. Tutak. 2023. "Credit Risk Modeling Using Interpreted XGBoost." *European Management Studies* 21, no. 3: 46–70.
- Hervani, A. A., M. M. Helms, and J. Sarkis. 2005. "Performance Measurement for Green Supply Chain Management." *Benchmarking: An International Journal* 12, no. 4: 330–353.
- Heyes, A. 2000. "Implementing Environmental Regulation: Enforcement and Compliance." *Journal of Regulatory Economics* 17, no. 2: 107–129.
- Karpoff, J. M., J. R. Lott, and G. Rankine. 1999. *Environmental Violations, Legal Penalties, and Reputation Costs*. University of Chicago, Law School.
- Khan, S. J., A. Dhir, V. Parida, and A. Papa. 2021. "Past, Present, and Future of Green Product Innovation." *Business Strategy and the Environment* 30, no. 8: 4081–4106.
- King, A. A., and M. J. Lenox. 2000. "Industry Self-Regulation Without Sanctions: The Chemical Industry's Responsible Care Program." *Academy of Management Journal* 43, no. 4: 698–716.
- Kirk, E. A., A. D. Reeves, and K. L. Blackstock. 2007. "Path Dependency and the Implementation of Environmental Regulation." *Environment and Planning. C, Government & Policy* 25, no. 2: 250–268.
- Konar, S., and M. A. Cohen. 2001. "Does the Market Value Environmental Performance?" *Review of Economics and Statistics* 83, no. 2: 281–289.

- Kumari, P., A. Shankar, R. Alghafes, L. Broccardo, and S. T. Walsh. 2025. "Exploring the Interrelationship Between Energy, Geopolitical Risk, and Bitcoin Based Green Business Strategies." *Business Strategy and the Environment* 34: 6654–6669.
- Laplace, B., and P. Lanoie. 1994. "The Market Response to Environmental Incidents in Canada: A Theoretical and Empirical Analysis." *Southern Economic Journal* 60: 657–672.
- Leone, A. J., M. Minutti-Meza, and C. E. Wasley. 2019. "Influential Observations and Inference in Accounting Research." *Accounting Review* 94, no. 6: 337–364.
- Mangalathu, S., S. H. Hwang, and J. S. Jeon. 2020. "Failure Mode and Effects Analysis of RC Members Based on Machine-Learning-Based SHapley Additive exPlanations (SHAP) Approach." *Engineering Structures* 219: 110927.
- Maqbool, M., and I. H. Qureshi. 2026. "A Systematic Literature Review on Explainable AI for Churn Prediction, Customer Segmentation and Retention." *International Journal of Data Science and Analytics* 22, no. 1: 32.
- Marinovic, I., and F. Varas. 2019. "CEO Horizon, Optimal Pay Duration, and the Escalation of Short-Termism." *Journal of Finance* 74, no. 4: 2011–2053.
- Nandha Gopal, J., and S. Muthukaruppasamy. 2024. "Precision Poverty Evaluation: Leveraging Generalized Linear Models and XG Boost Algorithms." In *Smart Technologies for Sustainable Development Goals: No Poverty*, 214. CRC Press.
- Nsir, N., A. B. Alzubi, and O. R. Adegboye. 2025. "Enhancing Sustainable Supply Chain Performance Prediction Using an Augmented Algorithm-Optimized XGBOOST in Industry 4.0 Contexts." *Sustainability* 17, no. 22: 10344.
- Pahari, S., D. Chakraborty, A. Polisetty, G. Dash, M. A. Camilleri, and J. Zhang. 2024. "Factors Affecting Consumer Purchases of Natural Foods: Prioritizing Health Consciousness and Environmental Sustainability." *Business Strategy and the Environment* 33, no. 8: 8247–8266.
- Pensa, R. G., A. Crombach, S. Peignier, and C. Rigotti. 2025. "Explaining Random Forest and XGBoost With Shallow Decision Trees by Co-Clustering Feature Importance." *Machine Learning* 114, no. 12: 287.
- Potoski, M., and A. Prakash. 2005. "Covenants With Weak Swords: ISO 14001 and Facilities' Environmental Performance." *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 24, no. 4: 745–769.
- Qin, C., Y. Zhang, F. Bao, C. Zhang, P. Liu, and P. Liu. 2021. "XGBoost Optimized by Adaptive Particle Swarm Optimization for Credit Scoring." *Mathematical Problems in Engineering* 2021, no. 1: 1–18.
- Sharma, S., and I. Henriques. 2005. "Stakeholder Influences on Sustainability Practices in the Canadian Forest Products Industry." *Strategic Management Journal* 26, no. 2: 159–180.
- Shen, H., Z. Cao, and J. Liu. 2025. "Administrative Decentralization, Government-Firm Interaction and Firm Pollution Emissions." *Journal of Cleaner Production* 519: 145952.
- Shevchenko, A. 2021. "Do Financial Penalties for Environmental Violations Facilitate Improvements in Corporate Environmental Performance? An Empirical Investigation." *Business Strategy and the Environment* 30, no. 4: 1723–1734.
- Shimshack, J. P., and M. B. Ward. 2008. "Enforcement and Over-Compliance." *Journal of Environmental Economics and Management* 55, no. 1: 90–105.
- Simsek, A. I., E. Koc, and E. Gultekin Tarla. 2025. "Assessing Green Logistics and Supply Chain Resilience With Future Importance Analysis: Machine Learning and Multicriteria Decision-Making Approach." *Business Strategy and the Environment* 35: 3952–3977.
- Singh, A., S. S. Almasabi, A. K. Patel, and P. Malik. 2025. "Sailing From Penalties to Accountability: Business Strategies and Governance for Firms to Innovate After Environmental Misconduct." *Business Strategy and the Environment* 35: 6382–6398.
- Sørensen, J. B., and T. E. Stuart. 2000. "Aging, Obsolescence, and Organizational Innovation." *Administrative Science Quarterly* 45, no. 1: 81–112.
- Testa, F., O. Boiral, and F. Iraldo. 2018. "Internalization of Environmental Practices and Institutional Complexity: Can Stakeholders Pressures Encourage Greenwashing?" *Journal of Business Ethics* 147, no. 2: 287–307.
- Thornton, D., R. A. Kagan, and N. Gunningham. 2008. "Compliance Costs, Regulation, and Environmental Performance: Controlling Truck Emissions in the US." *Regulation & Governance* 2, no. 3: 275–292.
- Tilley, F. 2000. "Small Firm Environmental Ethics: How Deep Do They Go?" *Business Ethics: A European Review* 9, no. 1: 31–41.
- Treepongkaruna, S., H. H. Au Yong, K. Kyaw, and M. R. Kalhor. 2025. "Biodiversity Reporting and Climate Policy Uncertainty." *Business Strategy and the Environment* 35: 1060–1075.
- Vogel, D. 2007. *The Market for Virtue: The Potential and Limits of Corporate Social Responsibility*. Bloomsbury Publishing USA.
- Walden, W. D., and A. J. Stagliano. 2003. "An Assessment of the Quality of Environmental Disclosure Themes." In *Advances in Environmental Accounting & Management*, 137–165. Emerald Group Publishing Limited.
- Wang, C., C. Deng, and S. Wang. 2020. "Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Binary Label-Imbalanced Classification With XGBoost." *Pattern Recognition Letters* 136: 190–197.
- Yi, Y., and P. Demirel. 2023. "The Impact of Sustainability-Oriented Dynamic Capabilities on Firm Growth: Investigating the Green Supply Chain Management and Green Political Capabilities." *Business Strategy and the Environment* 32, no. 8: 5873–5888.
- Yu, J., C. W. H. Lo, and P. H. Y. Li. 2017. "Organizational Visibility, Stakeholder Environmental Pressure and Corporate Environmental Responsiveness in China." *Business Strategy and the Environment* 26, no. 3: 371–384.
- Zhang, P., Y. Jia, and Y. Zhang. 2022. "Research and Application of XGBoost in Imbalanced Data." *International Journal of Distributed Sensor Networks* 18, no. 6: 155013292211069.
- Zhang, Y., and A. Haghani. 2015. "A Gradient Boosting Method to Improve Travel Time Prediction." *Transportation Research Part C: Emerging Technologies* 58: 308–324.
- Zhang, Y., X. Shi, S. Zhang, and A. Abraham. 2022. "A XGBoost-Based Lane Change Prediction on Time Series Data Using Feature Engineering for Autopilot Vehicles." *IEEE Transactions on Intelligent Transportation Systems* 23, no. 10: 19187–19200.
- Zou, H. L., R. C. Zeng, S. X. Zeng, and J. J. Shi. 2015. "How Do Environmental Violation Events Harm Corporate Reputation?" *Business Strategy and the Environment* 24, no. 8: 836–854.