Babylonian Journal of Machine Learning
Vol.2026, **pp**. 258–273
DOI: https://doi.org/10.58496/BJML/2026/016; ISSN: 3006-5429
https://mesopotamian.press/journals/index.php/BJML

Research Article

Explainable Prediction of Supplier Sanction Risk in Public Procurement: Comparing Random Forest and XGBoost under Forward Temporal Validation

Miluska Odely Rodríguez Saavedra [1,*], Robinson Bernardino Almanza Cabe [1], Rubén Washington Arguedas Catasi [1]

[1] *Graduate School, Universidad Nacional de San Agustín de Arequipa, Arequipa, Peru.*

ARTICLE INFO

Article History
Received 22 Jul 2026
Revised 27 Aug 2026
Accepted 25 Sep 2026
Published 02 Nov 2026

Keywords
Public procurement,
Supplier sanctions,
Random Forest,
XGBoost,
SHAP,
Temporal validation,
Sanction registry.

ABSTRACT

Administrative sanctions against suppliers to the Peruvian state are imposed after the events that give rise to them, which limits their usefulness for guiding anticipatory oversight. This study evaluates the ability of Random Forest and XGBoost to predict debarments or fines in the following year based on the contractual and sanction trajectory of suppliers. A panel of 210,434 supplier-year observations from 94,359 suppliers awarded contracts between 2018 and 2025 was built, and the sanction history was reconstructed from archived versions of administrative registries. Using forward temporal validation over 2022–2024, both algorithms were compared with logistic regression and a model based only on prior sanctions. Random Forest achieved the best performance, with a ROC-AUC of 0.792 and a capture of 42.7% of sanctioned suppliers when the 10% with the highest risk were reviewed, compared with 0.789 and 41.5% for XGBoost. The difference between the two was significant only in PR-AUC. Annual performance ranged from 0.766 to 0.808 and reached 0.769 among suppliers without prior sanctions, who accounted for 75.2% of sanctioned suppliers. The temporal audit of sanctions and penalties barely changed performance, whereas using the current registry instead of the reconstructed one reduced the ROC-AUC by 0.027. With Random Forest, 12.5 suppliers had to be reviewed for each sanctioned supplier identified, compared with 53.2 under random selection. This study demonstrates that the reconstruction of the sanction history and the application of Random Forest and XGBoost make it possible to anticipate the risk of administrative sanction and to direct oversight toward the suppliers with the highest risk.

1. INTRODUCTION

Public procurement is the channel through which the state acquires works, goods, and services from private suppliers, and its weight in public spending makes the reliability of those suppliers a condition for effective government. The outcome of each contract depends both on the rules that govern supplier selection and on the capacity of contracting entities to oversee those who deliver the work [1]. When a contractor breaches its obligations, submits false information, or abandons execution, the cost falls on stalled works, interrupted services, and funds that must be committed again. Award data make it possible to build objective risk signals before any formal investigation exists [2], but these signals describe procedures rather than the accumulated conduct of each supplier over time.

Administrative sanctions are the system's response to such breaches. Debarment excludes a supplier from the public market for a period and serves as a protective mechanism for contracting entities [3], and its preventive use against corruption has been extended to the exclusion of convicted bidders [4]. Its limitation is timing, because the sanction arrives after the facts that motivate it and only after a procedure subject to deadlines and statutes of limitation [5]. During that interval the supplier remains eligible to win contracts, so a sanction that is effective as punishment arrives too late to prevent the damage.

Anticipating sanction risk moves oversight to the stage at which it is still useful. This type of task belongs to the policy decisions that rely on a prediction, such as choosing where to concentrate a limited inspection capacity [6]. Analytical tools for public procurement have focused mainly on collusion among bidders, fraud signals, and anomalies in individual contracts [7], while the subsequent sanction of the supplier is rarely used as the outcome to be anticipated.

The Peruvian case offers the conditions to examine this problem and adds a specific difficulty. The Public Procurement Tribunal records thousands of debarments and fines every year, but the public registry of sanctioned suppliers removes sanctions that expired more than four years earlier. Of the 1,681 debarments that started in 2019, the current registry retains only 420, so a model built on that registry underestimates supplier histories and the frequency of sanctions in the years used

*Corresponding author. Email: mrodriguezsaa@unsa.edu.pe

for training. This study therefore addresses the following research question. To what extent does the contractual and sanction trajectory of a supplier, observed at the end of a year, predict its sanction by the Public Procurement Tribunal in the following year?

The general objective is to evaluate the explainable prediction of supplier sanction risk in public procurement through the comparison of Random Forest and XGBoost with forward temporal validation. The specific objectives are (i) to compare the predictive performance of Random Forest, XGBoost, multivariable logistic regression, and logistic regression based on the prior sanctions of suppliers to the state; (ii) to analyze the predictive performance of Random Forest and XGBoost among suppliers to the Peruvian state without prior sanctions; (iii) to identify through SHAP the variable groups with the largest contribution to the Random Forest predictions and to contrast their importance with XGBoost; (iv) to estimate the effect of sanction registry completeness on the predictive capacity of the models by comparing the current registry with the reconstructed registry; and (v) to evaluate the practical usefulness of the predictions for prioritizing the oversight of suppliers with the highest sanction risk. The contribution of the study comprises a data component, based on the reconstruction of the historical sanction registry, and a methodological component, supported by forward temporal validation, the audit of predictor availability, the comparison of models, and the evaluation of their operational usefulness, while the algorithms used correspond to existing methods.

2. RELATED WORK

Becker [8] argues that the deterrent effect of a sanction depends on the probability of detection and on the severity of the expected punishment. In public procurement, Spagnolo [9] shows that supplier reputation can incorporate past performance into future awards without restricting competition, and Butler et al. [10] find that past-performance mechanisms improve execution quality without substantially reducing the entry of new bidders. Gerardino et al. [11] document that agencies shift purchases toward less scrutinized procedures when the probability of an audit increases, so the recorded frequency of sanctions reflects both supplier conduct and the intensity of oversight.

Procedure design also conditions procurement outcomes. Coviello et al. [12] show that discretion favors repeated awards to local suppliers, while Palguta and Pertold [13] identify the bunching of contracts just below the thresholds that allow discretionary procedures. These results show that procurement characteristics and the relationship between supplier and entity contain signals that can be incorporated into risk models.

On this basis, Table I brings together related work on prediction and prioritization using administrative data. It includes studies on early warning, risk indicators, and the detection of collusion, anomalies, fraud, and corruption, which make it possible to situate the analysis of future administrative sanctions at the supplier-year level.

TABLE I. RELATED WORK ON RISK PREDICTION IN PUBLIC PROCUREMENT

Author(s) and year	Method or technique	Application	Main contribution	Difference from this study
Gallego et al. (2021) [14]	Early warning model with machine learning	Public procurement	Predicts the risk of misconduct with administrative records	Does not predict future administrative sanctions of the supplier
Fazekas and Kocsis (2020) [15]	Objective risk indicators	Public procurement in several countries	Builds corruption indicators from procurement patterns	Does not estimate future sanctions or use the supplier-year unit
Decarolis and Giorgiantonio (2022) [16]	Red flags and data analysis	Calls for tenders in Italy	Relates procurement alerts to subsequent investigations	The unit is the call for tenders, not the sanction of the supplier
Huber and Imhof (2019) [17]	Machine learning with statistical screens of bids	Collusion in Swiss public works tenders	Classifies competitive and collusive tenders from the distribution of bids	The unit is the tender and the outcome is collusion
García Rodríguez et al. (2022) [18]	Comparison of classification algorithms	Collusion in public auctions from several countries	Identifies the best-performing algorithms for detecting collusive patterns	Does not analyze the subsequent sanction of the supplier
Domingos et al. (2016) [19]	Deep learning for anomaly detection	IT purchases by the Brazilian federal government	Detects anomalous purchases as signals of potential fraud	Does not use an observed administrative sanction as the outcome
Velasco et al. (2021) [20]	Decision support system with risk indicators	Fraud in Brazilian public procurement	Ranks companies by fraud signals to guide investigations	Does not evaluate predictions on later years
López-Iturriaga and Sanz (2018) [21]	Neural networks	Public corruption in Spanish provinces	Generates alerts of corruption cases using economic and political variables	The unit is the province, not the supplier
Ash et al. (2025) [22]	Machine learning with administrative data	Municipal audits in Brazil	Prioritizes municipalities with higher corruption risk for auditing	The unit is the municipality and the outcome is audit findings

Author(s) and year	Method or technique	Application	Main contribution	Difference from this study
This study	Random Forest, XGBoost, logistic regression, and SHAP	Suppliers to the Peruvian state	Predicts debarment or fine in t+1 with a reconstructed sanction history	Forward temporal validation at the supplier-year level

Table I summarizes research on collusion, fraud, corruption, purchase anomalies, and audit prioritization. Their units of analysis include tenders, purchases, provinces, municipalities, and firms. In this study, the unit of analysis is the supplier-year and the outcome corresponds to a future administrative sanction; the design incorporates forward temporal validation and the comparison between the current registry and the reconstructed registry.

Breiman [23] proposed Random Forest as an ensemble of trees built on bootstrap samples and random subsets of predictors, and Chen and Guestrin [24] developed XGBoost as a regularized and scalable gradient boosting system for tabular data; both can represent nonlinear interactions among supplier scale, prices, and trajectory, and Shwartz-Ziv and Armon [25] show that tree ensembles match or outperform more costly deep architectures on tabular data. Lundberg et al. [26] formulated the exact computation of SHAP values for tree models, which distributes each prediction among the predictors and aggregates them into global explanations, whereas Rudin [27] cautions that, in high-stakes decisions, post hoc explanations of a complex model should be examined carefully and compared with interpretable alternatives. Moreno-Torres et al. [28] show that changes in the data distribution between training and deployment degrade performance, a condition to be expected in public procurement because of regulatory reforms and changes in oversight. In the reviewed work, supplier sanctions are rarely used as an outcome, evaluations on years after the training period are limited, and how the loss of historical sanctions from the registry affects model training is not examined.

3. METHODOLOGY

The procedure integrated award, consortium, penalty, and sanction records into a supplier-year panel, defined the sanction in t+1, built predictors with information available at the end of t, and evaluated four models through forward temporal validation, with SHAP explanations, an audit of the temporal availability of the predictors, and an analysis of sanction registry completeness. Fig. 1 depicts the complete procedure.

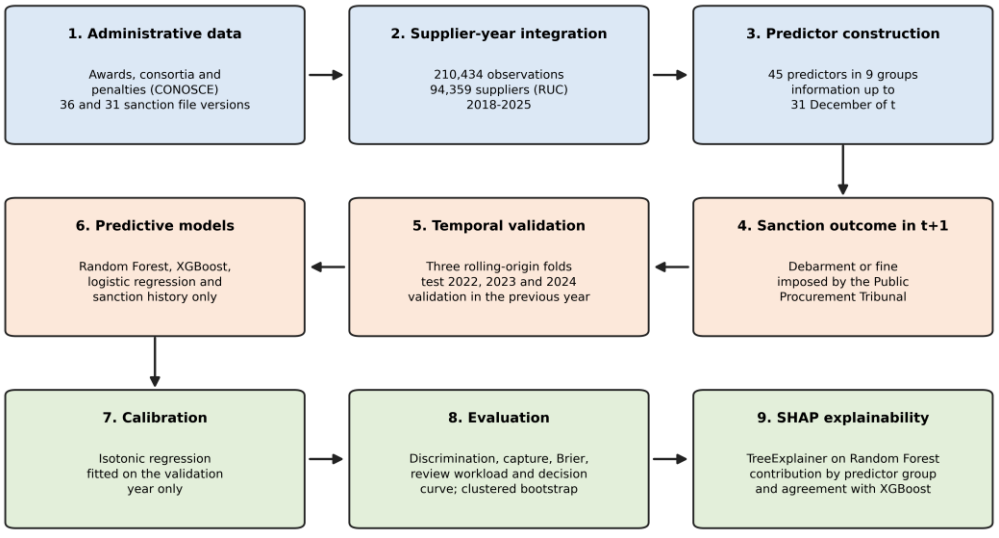


Fig. 1. Methodological workflow of the study, from the integration of administrative data to the explanation of predictions with SHAP.

The workflow comprises nine stages. The first four stages build the supplier-year panel from the administrative records, compute the predictors with information available at the end of each year, and define the sanction outcome in the following year. Stages five to eight organize the forward temporal validation, estimate and recalibrate the four models, and evaluate them, and the last stage explains the Random Forest predictions with SHAP. The following subsections describe each stage.

3.1 Dataset Description

The population consists of the suppliers that won at least one award in Peruvian public procurement between 2018 and 2025, including members of winning consortia, according to the open data of the Agency for Efficient Public Procurement (OECE), formerly the Supervisory Agency for State Procurement. The data come from the award, direct contracting, and consortium files of the CONOSCE system, from the registries of suppliers sanctioned with debarment and with fines, and from the registry of contractual penalties, downloaded on September 29, 2026, with the sanction registry updated to September 26 of that year. The unit of analysis is the supplier, identified by its taxpayer number (RUC), in each year t in which it won at least one award.

The panel contains 210,434 observations from 94,359 suppliers. Observations from 2025 were excluded from evaluation because their outcome covers only sanctions that started up to September 26, 2026, so the predictive analysis uses 181,032 observations from 2018 to 2024. The annual share of suppliers sanctioned in the following year ranged from 1.6% in 2024 to 3.1% in 2020, while the share of suppliers with prior sanctions increased from 1.4% in 2018 to 6.8% in 2024. For consortium awards, the amount was split equally among members because participation shares are not published.

The sanction history was reconstructed because the current registry removes sanctions that expired more than four years earlier. Merging 36 archived versions of the debarment file and 31 versions of the fines file recovered 16,635 debarments and 6,564 fines, compared with 9,855 and 4,151 retained in the current registry. The outcome variable and the history variables of the panel were independently reproduced from the reconstructed registry, with full agreement for the outcome and agreement above 99.9% for the number of prior sanctions. To audit temporal availability, the complete history of versions published between September 29, 2021, and October 10, 2026, downloaded in October 2026, was also used, with 354 versions of the debarment file, 310 of the fines file, and 301 and 294 of the two penalty files, whose upload date indicates when each record first became public. For sanctions that started from October 2021 onward, the median lag between the start date and the first publication was 3 days, with a 95th percentile of 17 days for debarments and 8 days for fines, whereas penalties had a median lag of 73 days and only between 54% and 59% of those dated in year t had been published by December 31 of t. No observable publication date exists for penalties recorded before September 2021.

3.2 Outcome Variable and Predictors

For each supplier i observed in year t, the outcome indicates the start of a debarment or a fine imposed by the Public Procurement Tribunal in the following year, according to the recorded start date, and is defined as

$$Y_{i,t+1} = \begin{cases} 1, & \text{if supplier } i \text{ starts a debarment or fine in } t+1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Debarments and fines were also analyzed separately as secondary outcomes. The 45 predictors, equivalent to 47 model inputs after encoding supplier type as indicator variables, are organized into nine analytical groups derived from the public procurement literature and were computed with information dated up to December 31 of year t, a necessary condition because information leakage occurs when the model uses information that was not yet available at the prediction date [29]. Table II presents the groups and their indicators.

TABLE II. PREDICTOR GROUPS AND INDICATORS

Predictor group	Indicators	Scale or source
Sanction history	Prior debarments and fines, years since the last one, and active sanction at the end of t	Registry reconstructed since 1998
Contract compliance	Contractual penalties in t and cumulative, and cumulative penalty amount	Penalty registry
Contracting scale	Items, procedures, awarded and reference amounts, entities, and departments in t	Logarithm for amounts and counts
Award price	Awarded/reference ratio, median discount, and share of items awarded at the reference value	Discount bounded between −1 and 1
Selection procedure	Share of items by direct contracting, emergency, reverse auction, simplified award, public tender, and price comparison; days from call to award; share awarded in December	Shares of awarded items
Dependence on entities	Share of the amount from the main entity, HHI across entities, and share with local and regional governments	HHI between 0 and 1
Contract category	Share of the amount in works, goods, services, and works consultancy services	Shares of the amount
Supplier trajectory	Years with awards since 2018, years since the first award, cumulative amount and items, award, amount, and entities in t-1, and growth in the amount	History observed since 2018
Form of participation	Share of the amount through consortia, number of consortia, and supplier type	Natural person, legal entity, or nonresident

Note. Indicators were built with records dated up to the end of each year. For sanctions and penalties, the date of first publication was also verified. HHI: Herfindahl-Hirschman index.

The nine groups in Table II cover the sanction and compliance record of the supplier, the scale and price of its awards, the procedures through which it obtains contracts, its dependence on particular buyers, the object of its contracts, its trajectory since 2018, and the form in which it participates. All of them are measured at the end of year t, so they approximate the information that a contracting entity or the regulator would have at the time of prediction.

Amounts and counts were transformed with log(1+x). For logistic regression, continuous variables were standardized with the mean and standard deviation of the training set of each fold as

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{2}$$

where x_{ij} is the value of predictor j for observation i, and μ_j and σ_j are its mean and standard deviation. XGBoost handled missing values natively, whereas for logistic regression and Random Forest they were imputed with values that preserve their meaning, such as 30 years since the last sanction for suppliers without prior sanctions or a growth ratio of 1 for suppliers without awards in t-1. Missing values in year-t indicators did not exceed 0.2%.

3.3 Temporal Validation

The evaluation was organized into three rolling-origin folds, in which each model is trained on earlier years, tuned on the year immediately before the test year, and evaluated on a year that takes no part in any modeling decision. This design reproduces the conditions under which the alert would be used, a requirement of out-of-sample evaluation to avoid overstating performance [30]. In the 2022 fold, training covers 2018 to 2020 with 71,637 observations, validation covers 2021, and testing covers 2022; the 2023 and 2024 folds extend the training period by one year each. The three test years contain 85,341 observations from 54,806 suppliers and 1,604 sanctions, for a rate of 1.88%. Test-year data were not used in the training, tuning, or calibration of the models.

3.4 Predictive Models

Four models were estimated on the same data in each fold. The simplest benchmark is a logistic regression with the six sanction-history indicators, which represents the rule of an entity that looks only at prior sanctions. The other three models use all 45 predictors. Logistic regression with L2 regularization estimates the probability of sanction as

$$P\left(Y_{i,t+1} = 1 \mid x_{it}\right) = \frac{1}{1+\exp\left(-\left(\beta_0 + x_{it}^{\top}\beta\right)\right)} \tag{3}$$

where x_{it} is the vector of standardized predictors and β_0 and β are the coefficients. Random Forest used 500 trees and a minimum of 20 observations per leaf, following the principle of aggregating trees built on bootstrap samples and random subsets of predictors. XGBoost builds an additive function of K regression trees of the form

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \; f_k \in \mathcal{F} \tag{4}$$

where each f_k belongs to the space F of regression trees. Trees are added sequentially by minimizing a regularized objective function

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

where l is the log loss, $\Omega(f) = \gamma T + \tfrac{1}{2}\lambda\|w\|^2$ penalizes the number of leaves T and the magnitude of their weights w, and γ and λ are regularization parameters. Maximum depths of 3, 5, and 7 and minimum child weights of 1 and 10 were evaluated, with a learning rate of 0.03, row and column subsampling of 0.8, and early stopping on the validation log loss with a patience of 200 iterations. The configuration with the lowest validation loss was a depth of 7 with 71 trees in the 2022 fold, a depth of 3 with 640 trees in the 2023 fold, and a depth of 5 with 180 trees in the 2024 fold. Classes were not reweighted, and the probabilities of all four models were recalibrated with isotonic regression fitted, in each fold, only on the scores and observed outcomes of the validation year; the fitted function was applied unchanged to the test year, a procedure that corrects the probability scale without altering the ranking [31].

3.5 Evaluation

Discrimination was measured with the area under the ROC curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC), the latter being more informative when the positive class is a minority and having a no-skill value equal to the sanction rate [32]. Both were computed on the original scores, because isotonic recalibration creates ties. Operational usefulness was measured by ranking suppliers by estimated risk and selecting the 10% with the highest scores, with capture and precision defined as

$$\text{Capture@10\%} = \frac{\text{TP}_{10}}{\text{TP}_{10}+\text{FN}_{10}} \tag{6}$$

$$\text{Precision@10\%} = \frac{\text{TP}_{10}}{\text{TP}_{10}+\text{FP}_{10}} \tag{7}$$

where TP_{10} are the suppliers sanctioned in t+1 within the top 10% of risk, FN_{10} are the sanctioned suppliers outside that group, and FP_{10} are the non-sanctioned suppliers within it. Review workload was expressed as the number of suppliers that must be reviewed for each sanctioned supplier identified, equal to 1/Precision@k for k of 5%, 10%, and 20%, and was compared with random selection. Decision curve analysis [33] compared the net benefit of each model, $NB = TP/N - (FP/N)\cdot p_t/(1 - p_t)$, where p_t represents the minimum estimated probability of sanction from which reviewing a supplier is considered justified, against the strategies of reviewing all suppliers or none. The quality of the recalibrated probabilities was summarized with the Brier score

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2 \tag{8}$$

where $\hat{p}_i$ is the recalibrated probability and N is the number of observations, together with a comparison between the mean predicted rate and the observed rate. The probabilities used in the Brier score and in the decision curve come exclusively from the recalibration fitted on the validation year. Metrics were computed for the pooled test years and for each year. The 95% confidence intervals were obtained with 1,000 bootstrap resamples clustered by supplier, which draw suppliers with replacement together with all their observations [34], and differences between models were assessed in a paired manner on the same resamples.

3.6 SHAP Explainability, Temporal Audit, and Registry Completeness

Predictor contributions were estimated with TreeExplainer on the Random Forest model of each fold and its test year, on the probability scale and before recalibration, using a random sample of 2,000 observations per test year; the same procedure was applied to XGBoost, on the logit scale, to contrast both rankings. The prediction for each observation is decomposed as

$$f(x) = \varphi_0 + \sum_{j=1}^{M}\varphi_j \tag{9}$$

where φ_0 is the base value of the model, φ_j is the contribution of predictor j, and M is the number of predictors. The contribution of each group was computed as the mean absolute value of the sum of the φ_j of its indicators, and the stability of the group ranking across folds was summarized with Spearman correlation. SHAP values were interpreted as contributions to the prediction and not as causal effects, because dependencies among predictors distort the causal reading of explanations [35]. The temporal audit recomputed performance by counting each sanction and each penalty only if it had been published before December 31 of t, first only in the test years and then also in the training years; sanctions present in the first archived version were dated with the 95th percentile of the observed lag. To estimate the effect of registry completeness, the entire procedure was repeated in four configurations that use the current or the reconstructed registry to define the training outcome and to compute the history indicators: A, both with the current registry; B, current outcome and reconstructed history; C, reconstructed outcome and current history; and D, both reconstructed, which corresponds to the main model. The four configurations were evaluated against the reconstructed outcome of the test years.

4. RESULTS

4.1 Comparative Performance of the Models

Table III summarizes the performance of the four models on the 85,341 test supplier-years, with a sanction rate of 1.88%. The three models that use all predictors performed better than the model based only on sanction history, while the differences among them were small.

TABLE III. PREDICTIVE PERFORMANCE IN THE 2022-2024 TEST YEARS

Model	ROC-AUC (95% CI)	PR-AUC (95% CI)	Capture@10% (%) (95% CI)	Precision@10% (%)	Brier
Sanction history only	0.623 (0.608-0.638)	0.055 (0.047-0.066)	26.9 (24.5-29.2)	5.1	0.0181
Logistic regression	0.774 (0.763-0.786)	0.085 (0.075-0.098)	41.0 (38.3-43.2)	7.7	0.0178
Random Forest	**0.792 (0.781-0.803)**	**0.087 (0.077-0.099)**	**42.7 (40.3-45.1)**	**8.0**	**0.0178**
XGBoost	0.789 (0.777-0.799)	0.081 (0.072-0.093)	41.5 (39.1-43.9)	7.8	0.0179
Paired differences					
Random Forest − history only	0.168 (0.153 to 0.185)	0.032 (0.022 to 0.041)	15.8 (13.5 to 18.2)		−0.0003
Random Forest − logistic regression	0.018 (0.012 to 0.024)	0.001 (−0.005 to 0.007)	1.8 (−0.1 to 3.7)		0.0000
Random Forest − XGBoost	0.003 (−0.001 to 0.008)	0.005 (0.001 to 0.010)	1.2 (−0.3 to 2.7)		−0.0001

Note. N = 85,341 supplier-year observations from 54,806 suppliers, with 1,604 observations sanctioned in t+1. The sanction rate, 1.88%, is the reference value for PR-AUC. Intervals from 1,000 bootstrap resamples clustered by supplier. Capture differences are expressed in percentage points.

Random Forest achieved the best overall performance, with a ROC-AUC of 0.792 and a PR-AUC of 0.087, 4.6 times the sanction rate. When the 10% of suppliers with the highest risk were reviewed, it captured 42.7% of sanctioned suppliers with a precision of 8.0%, compared with 26.9% for the history-only model, an advantage of 15.8 percentage points. Relative to XGBoost, the advantage was small, as the paired interval excluded zero for PR-AUC but not for ROC-AUC or capture, and in four replications with different seeds Random Forest outperformed XGBoost by between 0.003 and 0.005 in ROC-AUC. Relative to logistic regression, Random Forest obtained a ROC-AUC higher by 0.018, with no significant differences in PR-AUC or capture. The Brier score was virtually identical across the three full models.

Fig. 2 shows the ROC and precision-recall curves of the four models over the pooled test years.

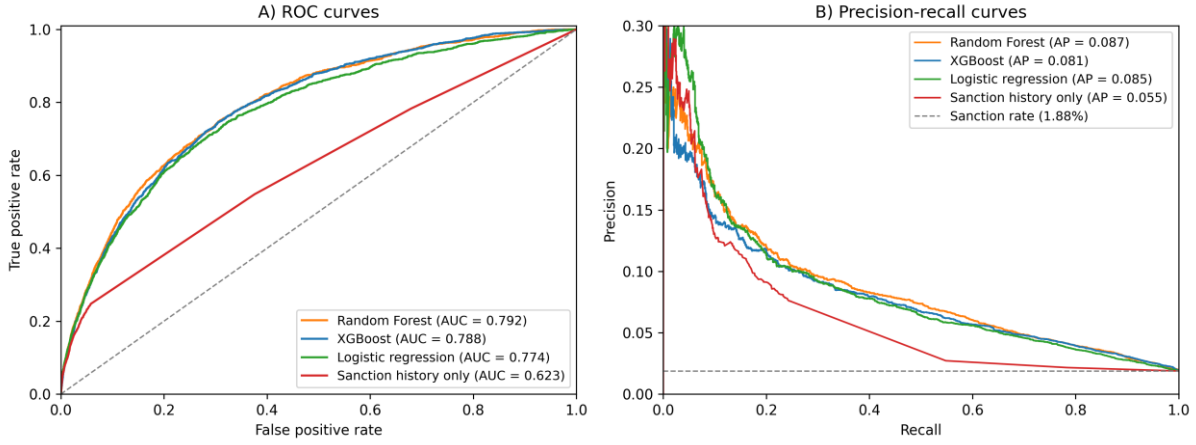


Fig. 2. ROC curves (A) and precision-recall curves (B) of the four models in the pooled 2022-2024 test years. AP denotes average precision, equal to the PR-AUC reported in Table III. The dashed line in (B) marks the sanction rate.

The ROC curves of Random Forest, XGBoost, and logistic regression overlap along the entire range, and at a false positive rate of 10% they identify 44.7%, 43.2%, and 41.9% of sanctioned suppliers, respectively, compared with 28.7% for the history-only model. At a false positive rate of 20%, the three full models reach between 60.9% and 62.7%, while the history-only model reaches 38.2%. The ROC curve of the history-only model rises steeply up to a true positive rate of about 0.24 at a false positive rate of 5.5% and then becomes almost linear, because most suppliers have no prior sanctions and receive the same score. In the precision-recall space, all models remain above the sanction rate at every level of recall, and Random Forest reaches a precision of 11.9% at a recall of 20% and 7.4% at a recall of 50%.

Fig. 3 presents the cumulative capture curves of the four models and the calibration of Random Forest by deciles of predicted probability in each test year.

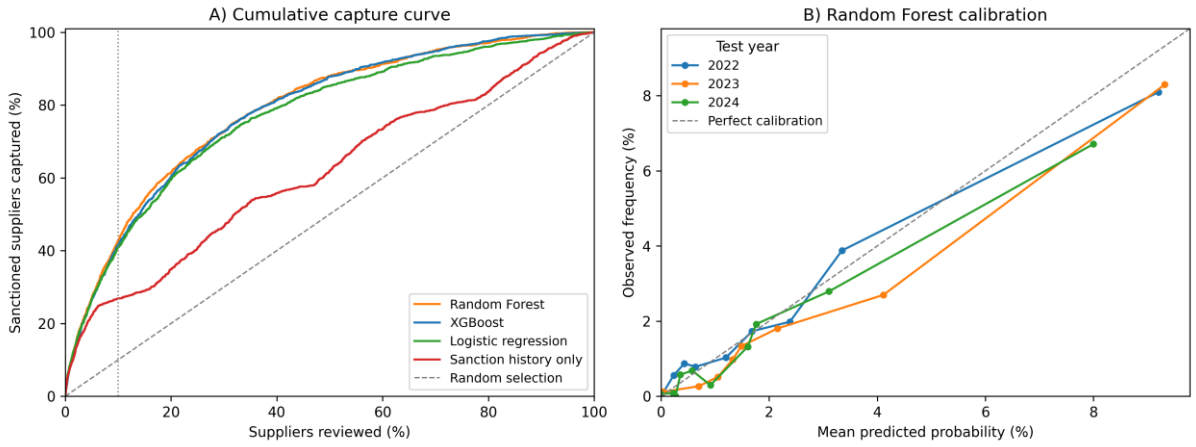


Fig. 3. Cumulative capture curve of the four models in the 2022-2024 test years (A) and calibration of Random Forest by deciles of recalibrated probability in each test year (B). The vertical dotted line marks 10% of suppliers reviewed.

Fig. 3A shows that the capture curves of the three full models overlap across the entire review range, whereas the curve of the history-only model flattens after the first 5% of suppliers, because beyond that point it can no longer distinguish among suppliers without prior sanctions. In Fig. 3B, the deciles lie close to the diagonal up to a predicted probability of about 4%, while the highest-risk decile falls below it in all three years, indicating that the model assigns probabilities above the observed frequencies among the suppliers with the highest estimated risk.

The review workload of Random Forest was 9.9, 12.5, and 17.3 suppliers for each sanctioned supplier identified when the 5%, 10%, and 20% with the highest risk were reviewed, compared with 12.0, 19.8, and 30.5 for the history-only model and 53.2 for random selection, while XGBoost and logistic regression required between 10.0 and 17.9. In the decision curve analysis, Random Forest obtained a higher net benefit than reviewing all or no suppliers for probability thresholds between 0.5% and 7.5%, with a maximum net benefit of 0.0078, equivalent to 7.8 net true positives per 1,000 suppliers evaluated, at the 2% threshold, and above 10% no model provided an appreciable net benefit.

Restricting the test-year predictors to the information published before each prediction date changed at least one predictor in 1.2% to 1.8% of the observations, mainly because of penalties not yet published, and did not appreciably change the performance of Random Forest, with a difference in ROC-AUC of −0.0003 (95% CI: −0.0007 to 0.0003) and in capture of 0.1 points (−0.5 to 0.6). A conservative re-estimation, which also excluded from training the records without an observable publication date, reduced the ROC-AUC of Random Forest to 0.784, with a difference of 0.008 (95% CI: 0.005 to 0.011), and capture to 40.5%, and kept the same ranking of the models. This reduction comes from the 2018–2020 training years, prior to the first archived version, and not from the evaluation years.

4.2 Performance by Test Year and Type of Sanction

Table IV breaks down the performance of Random Forest by test year and by type of sanction, together with the ROC-AUC of XGBoost and the history-only model for comparison.

TABLE IV. PERFORMANCE OF RANDOM FOREST BY TEST YEAR AND TYPE OF SANCTION

Evaluation	N	Sanctioned	Random Forest ROC-AUC (95% CI)	XGBoost ROC-AUC	History-only ROC-AUC	Random Forest PR-AUC	Random Forest Capture@10% (%)
2022	28,652	627	0.766 (0.747-0.784)	0.764	0.592	0.090	38.4
2023	27,534	503	0.808 (0.791-0.825)	0.797	0.598	0.101	46.3
2024	29,155	474	0.807 (0.790-0.826)	0.806	0.603	0.072	43.2
Debarment, 2022-2024	85,341	996	0.836 (0.825-0.848)	0.816	0.613	0.076	50.9
Fine, 2022-2024	85,341	657	0.723 (0.704-0.741)	0.714	0.492	0.027	32.1

Note. Debarment and fine were estimated separately and are not mutually exclusive categories; a supplier-year may record both. 95% CI from 1,000 bootstrap resamples clustered by supplier.

The ROC-AUC of Random Forest increased from 0.766 in 2022 to 0.808 in 2023 and 0.807 in 2024, in parallel with the extension of the training period, while the PR-AUC fell to 0.072 in 2024 as the sanction rate declined. XGBoost followed

the same trajectory with equal or slightly lower values in each year, and the history-only model remained close to 0.6 in all three years.

Calibration of Random Forest was adequate in 2022, with a mean predicted rate of 2.25% against an observed rate of 2.19%, but the model overestimated risk in 2023, with 2.37% against 1.83%, and in 2024, with 1.89% against 1.63%. Calibration slopes between 0.90 and 1.07 indicate that the miscalibration lies in the overall level of risk rather than in the spread of the probabilities, as Fig. 3B shows.

Debarment was more predictable than fines, with a ROC-AUC of 0.836 versus 0.723 for Random Forest and of 0.816 versus 0.714 for XGBoost, and a capture of 50.9% versus 32.1%. The history-only model reached 0.613 for debarment and 0.492 for fines, a value equivalent to a no-skill classifier. In the reconstructed history, 95.8% of the 6,249 fines that started between 2019 and 2025 correspond to failure to sign the contract after the award, while 60.4% of the 10,299 debarments in the same period originate in the submission of inaccurate information or false documents and 23.7% in contract termination by the contracting entity.

4.3 Suppliers With and Without Prior Sanctions

Table V reports the performance of the models by supplier history, in its upper block, and by the completeness of the registry used to train Random Forest, in its lower block.

TABLE V. PERFORMANCE BY SUPPLIER HISTORY AND SANCTION REGISTRY COMPLETENESS, 2022-2024

Scenario	N	Sanctioned	ROC-AUC (95% CI)	PR-AUC (95% CI)	Capture@10% (%) (95% CI)
Subgroups by history (reconstructed registry)					
Random Forest, all suppliers	85,341	1,604	0.792 (0.781-0.803)	0.087 (0.077-0.099)	42.7 (40.3-45.1)
Random Forest, no prior sanctions	80,082	1,207	0.769 (0.756-0.781)	0.055 (0.049-0.063)	38.9 (36.3-41.8)
XGBoost, no prior sanctions	80,082	1,207	0.769 (0.756-0.781)	0.051 (0.045-0.057)	37.4 (34.7-39.7)
History only, no prior sanctions	80,082	1,207	0.538 (0.522-0.553)	0.017 (0.015-0.018)	6.0 (4.8-7.3)
Random Forest, with prior sanctions	5,259	397	0.690 (0.663-0.718)	0.162 (0.136-0.201)	26.7 (22.4-30.9)
Registry completeness (Random Forest, evaluated against the reconstructed outcome)					
A. Outcome and history from the current registry	85,341	1,604	0.765 (0.753-0.777)	0.070 (0.063-0.079)	39.5 (37.2-41.8)
B. Current outcome and reconstructed history	85,341	1,604	0.776 (0.764-0.787)	0.079 (0.071-0.090)	41.2 (38.5-43.4)
C. Reconstructed outcome and current history	85,341	1,604	0.782 (0.771-0.793)	0.077 (0.069-0.087)	41.3 (38.8-43.6)
D. Reconstructed outcome and history (main model)	85,341	1,604	0.792 (0.781-0.803)	0.087 (0.077-0.099)	42.7 (40.3-45.1)
Effect of the reconstructed outcome (C − A)			**0.017 (0.012 to 0.023)**	**0.008 (0.003 to 0.013)**	**1.8 (0.1 to 3.7)**
Effect of the reconstructed history (B − A)			0.011 (0.007 to 0.014)	0.009 (0.006 to 0.014)	1.7 (0.3 to 2.8)
Total effect of completeness (D − A)			0.027 (0.021 to 0.034)	0.017 (0.011 to 0.024)	3.2 (1.1 to 5.3)

Note. The interval for the subgroup with prior sanctions was obtained with 200 resamples and the others with 1,000 bootstrap resamples clustered by supplier. Capture differences are expressed in percentage points.

Observations corresponding to suppliers without prior sanctions represent 93.8% of the test observations and account for 1,207 of the 1,604 cases sanctioned in t+1, 75.2% of the total, with a rate of 1.51% compared with 7.55% among suppliers already sanctioned. In this subgroup, Random Forest and XGBoost achieved a ROC-AUC of 0.769 and captured 38.9% and 37.4% of sanctioned suppliers, respectively, while the history-only model obtained 0.538. Among suppliers with prior sanctions, the ROC-AUC of Random Forest was 0.690.

4.4 SHAP Analysis

When the SHAP values of Random Forest were aggregated by group, supplier trajectory accounted for 24.9% of the total contribution, followed by contracting scale with 18.9% and sanction history with 11.1%. Dependence on entities contributed 10.9%, award price 10.0%, contract category 8.0%, selection procedure 6.2%, contract compliance 5.7%, and form of participation 4.3%. The group ranking was stable across folds, with Spearman correlations between 0.97 and 0.98. In XGBoost, trajectory also ranked first, with 32.0%, followed by dependence on entities with 14.1% and scale with 11.3%; the Spearman correlation between the group rankings of both models was 0.77, and 9 of the 12 most influential indicators coincided.

Fig. 4 shows the SHAP values of the twelve most influential indicators of Random Forest, ordered by their mean contribution.

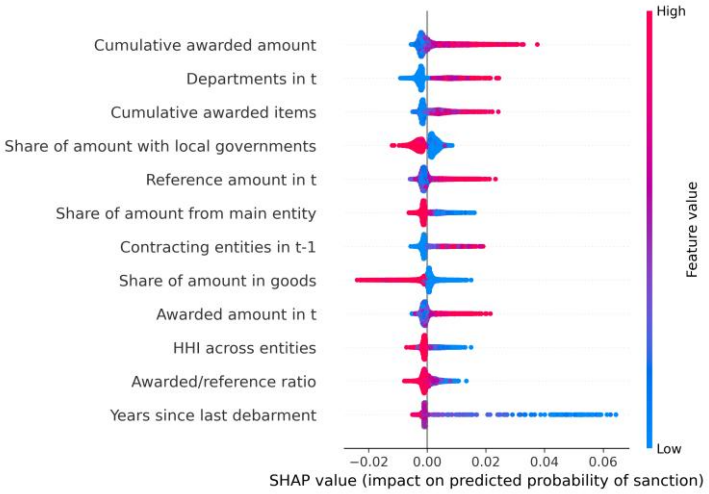


Fig. 4. SHAP values of the twelve indicators with the largest contribution to the Random Forest predictions in the 2022-2024 test years, based on a random sample of 6,000 observations (2,000 per year). Each point corresponds to one observation and its color indicates the value of the indicator; gray points in years since the last debarment correspond to suppliers without prior debarments.

In Fig. 4, the cumulative awarded amount, the number of departments in which the supplier executes contracts, the cumulative number of awarded items, the reference and awarded amounts in t, and the contracting entities in t-1 increase the estimated risk, whereas the share of the amount contracted with local governments, the share of the amount from the main entity, concentration across entities, and the share of the amount in goods reduce it. A lower awarded/reference ratio, that is, a larger discount, increases risk, and a recent debarment also increases it, with an effect that fades as years go by. Fig. 5 presents the Spearman correlation matrix of these twelve indicators in the test years.

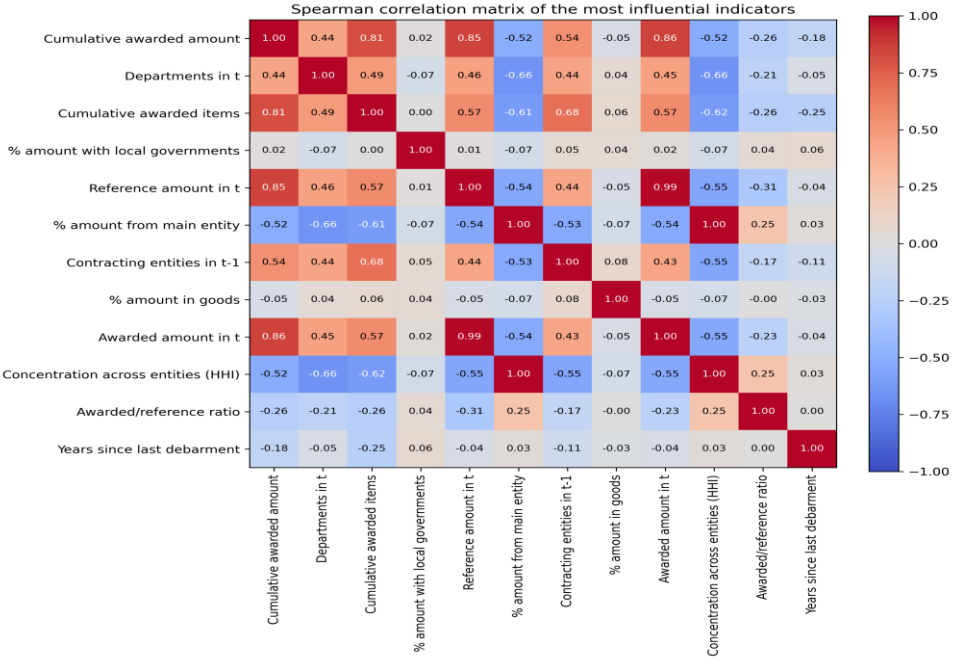


Fig. 5. Spearman correlation matrix of the twelve most influential indicators according to the Random Forest SHAP values, computed on the 2022-2024 test observations.

The median absolute correlation in Fig. 5 is 0.25, and 21 of the 66 pairs exceed 0.5 in absolute value. The share of the amount from the main entity and concentration across entities correlate at practically 1, the reference and awarded amounts in t at 0.99, the cumulative awarded amount with the awarded amount in t and with the reference amount at 0.86 and 0.85,

and with cumulative items at 0.81, while the number of departments is negatively related to concentration across entities and to the share of the main entity (−0.66). The strongest associations occur within the scale and dependence groups, and between scale and trajectory, which share part of their SHAP contribution, so the aggregation by group provides a more stable reading than the ranking of individual indicators within each group. The correlations involving years since the last debarment are computed only for the 4.5% of supplier-years with a prior debarment.

4.5 Reconstructed History Versus the Current Registry

The current registry omits 6,780 debarments and 2,413 fines, 40.8% and 36.8% of the reconstructed total, and the loss is concentrated in sanctions that started up to 2022. Fig. 6 compares the sanctions of the reconstructed history with those retained in the current registry and the resulting sanction rate in t+1.

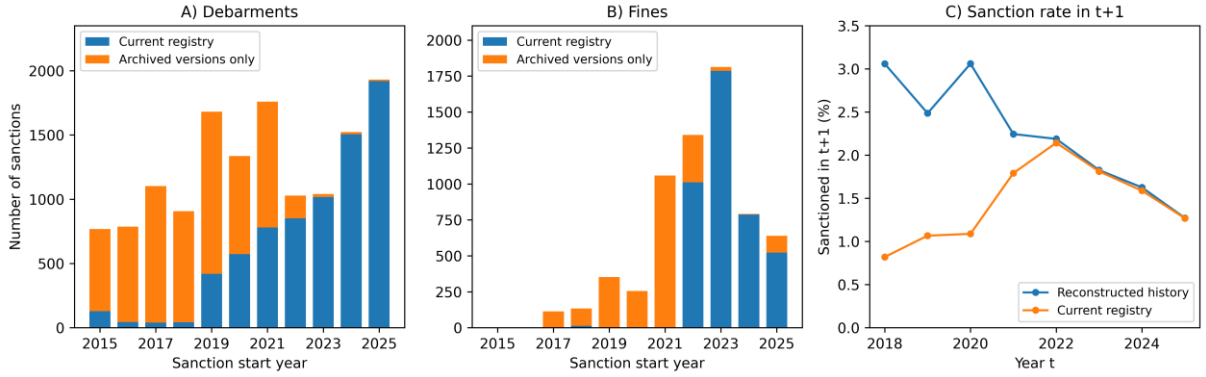


Fig. 6. Debarments (A) and fines (B) by start year according to their presence in the current registry or only in archived versions, and rate of suppliers sanctioned in t+1 computed with the reconstructed history and with the current registry (C).

Fig. 6A and Fig. 6B show that most debarments and fines that started before 2019 are missing from the current registry, and that the gap narrows sharply from 2022 onward. As a result, in Fig. 6C the sanction rate in t+1 computed with the current registry is 0.82% for 2018, compared with 3.06% with the reconstructed history, and that both rates coincide only from 2022 onward.

With the current registry, the training sets lose most of their positive cases, as the 2022 fold has 712 sanctioned suppliers instead of 2,058, the 2023 fold 1,143 instead of 2,598, and the 2024 fold 1,756 instead of 3,225, while the test years barely change, with 1,575 sanctioned suppliers versus 1,604. The lower block of Table V separates the two uses of the registry. With the history from the current registry, using the reconstructed outcome raised the ROC-AUC of Random Forest by 0.017 (C versus A), and with the current outcome, using the reconstructed history raised it by 0.011 (B versus A). Both effects were nearly additive and together amounted to 0.027 in ROC-AUC, 0.017 in PR-AUC, and 3.2 points in capture (D versus A), with intervals that exclude zero; XGBoost showed the same pattern, with a total gain of 0.026. The model trained on the current registry reports a ROC-AUC of 0.763 when evaluated on that same registry and 0.765 when evaluated on the reconstructed history, so the loss of information is not visible when only the current registry is analyzed, and the history-only model falls to a ROC-AUC of 0.494.

5. DISCUSSION

Random Forest anticipated next-year sanctions with a ROC-AUC of 0.792 and captured 42.7% of sanctioned suppliers within the top 10% of risk, with a small advantage over XGBoost and a capture close to that of logistic regression. This proximity among the models is consistent with Christodoulou et al. [36], who found no systematic advantage of machine learning over logistic regression in prediction models. The lower performance relative to the collusion studies summarized in Table I is explained by the fact that a sanction depends on breaches during execution and on the entity reporting them, not only on patterns observable in a tender. The result also shows that the usefulness of the alert comes from combining contractual and sanction information rather than from the complexity of the algorithm. Restricting the sanction and penalty predictors to the information published before each prediction date did not appreciably change performance, which supports the stability of the results against the publication lags identified in these sources. With a precision of 8.0% in the top 10% of risk, the tool reduces from 53.2 to 12.5 the number of suppliers that must be reviewed for each sanctioned supplier identified, although 11 of every 12 prioritized suppliers will not be sanctioned, so the list serves to order oversight and does not constitute evidence of liability.

Discrimination remained between 0.766 and 0.808 over the three test years, but calibration deteriorated when the sanction rate fell below that of the validation year. This pattern corresponds to the drift in the prior probability of the outcome

described by Gama et al. [37] and agrees with Van Calster et al. [38], who note that good discrimination does not guarantee correctly calibrated probabilities, so recalibration should be repeated when the frequency of the outcome changes. The difference between debarments and fines, with ROC-AUC values of 0.836 and 0.723, is consistent with Gallego et al. [39], who showed that predictive capacity changes with the type of outcome in administrative data. In this study, the outcome is a recorded administrative sanction and not the underlying conduct. The composition of the registry explains part of this difference, because 95.8% of fines penalize the failure to sign the contract after the award, a decision taken at the award stage that is unrelated to prior sanctions.

Supplier trajectory, contracting scale, and sanction history accounted for the largest contribution in Random Forest, and the order of the groups essentially matched that of XGBoost, while 75.2% of sanctioned suppliers had no prior sanctions, a group in which Random Forest retained a ROC-AUC of 0.769. The directions observed for award price agree with Calveras et al. [40] and Decarolis [41], because larger discounts increased risk, in line with the theory of reckless bids and their subsequent cost overruns. The higher risk associated with cumulative exposure to the public market is consistent with Colonnelli et al. [42], who link dependence on public contracts with the consequences firms face when they are sanctioned. The use of signals other than sanction history also coincides with the prioritization systems summarized in Table I. An alert based only on prior sanctions would miss most of the suppliers that will be sanctioned.

The completeness of the registry raised the ROC-AUC of Random Forest by 0.027 and capture by 3.2 percentage points, an improvement that comes mostly from recovering past sanctions as outcomes (0.017) and to a lesser extent from recovering them as supplier history (0.011). This improvement is not visible with the current registry, because the model trained on it reports 0.763 when its performance against the reconstructed outcome is 0.765. The problem resembles the selective labels problem described by Kleinberg et al. [43], in which an evaluation based on incompletely observed outcomes can conceal relevant differences, and the data construction errors that Kapoor and Narayanan [44] identify as a source of optimistic conclusions in applied machine learning. Hino et al. [45] showed that historical records can be used to target oversight tasks, and the results of this study indicate that such gains require preserving the complete history. Publishing all sanctions imposed by the Tribunal, not only recent ones, strengthens the construction and evaluation of alerts with the system's own data. More than three quarters of the debarments absent from the current registry entered the public file in a single historical load in November 2021, so the reconstruction depends on versions that the regulator no longer publishes [46-47].

6. LIMITATIONS

The limitations of the study relate to the selection of awarded suppliers and to the nature of administrative sanctions, whose imposition depends on detection and resolution procedures, so the predictions do not represent evidence of fraud or liability of the supplier. In addition, the historical availability of the records is subject to restrictions, because verifiable publication dates are not available for some penalties, nor are historical versions of awards and consortia. The temporal audit of sanctions and penalties showed that these restrictions did not substantially alter predictive performance in the evaluated years, although the results do not allow this conclusion to be extended to the sources without a publication history.

7. CONCLUSIONS

The information available at the end of each year made it possible to anticipate the administrative sanction of suppliers to the state in the following year. Random Forest achieved the best overall performance, with a ROC-AUC of 0.792 and 42.7% of sanctioned suppliers concentrated within the 10% with the highest estimated risk, with a small advantage over XGBoost, which reached 0.789, and a capture close to that of logistic regression. Restricting the sanction and penalty predictors to information published before each prediction date did not appreciably change model performance. Discrimination remained between 0.766 and 0.808 during 2022-2024, although calibration required adjustment in response to annual variations in the frequency of sanctions. Debarment showed greater predictability than fines, with a ROC-AUC of 0.836 versus 0.723.

Supplier trajectory was the predictor group with the largest contribution in Random Forest, followed by contracting scale and sanction history, an order that essentially matches XGBoost. Sanction history, however, proved insufficient as a sole alert criterion. Of the subsequently sanctioned suppliers, 75.2% had no prior sanctions, and even within this group Random Forest retained a ROC-AUC of 0.769. These findings show that early identification of sanction risk requires integrating the contractual trajectory of the supplier with different signals of its participation in public procurement, instead of concentrating oversight only on those who have already been sanctioned.

The reconstruction of the sanction history showed that the completeness of the records directly influences predictive performance. The current registry omits a substantial share of older sanctions, and the model trained on that information lost 0.027 in ROC-AUC, 0.017 in PR-AUC, and 3.2 percentage points in capture at 10% compared with the model built on the reconstructed history, a loss that comes mainly from the outcome and to a lesser extent from the history. This loss is difficult

to identify when the model is evaluated only against the current registry itself. Preserving the complete sanction history strengthens the training, evaluation, and reproducibility of alert systems. Within the top 10% of risk, the model requires reviewing 12.5 suppliers for each sanctioned supplier identified, compared with 53.2 under random selection, and it should be used to prioritize oversight actions and not to impose sanctions, restrictions, or administrative decisions automatically, because it predicts a future administrative sanction and not fraud, corruption, or liability of the supplier.

Data and Code Availability

The administrative records underlying this study are publicly available from Peru's public procurement authority. Additional information regarding the analytical dataset and research procedures may be obtained from the corresponding author upon reasonable request at mrodriguezsaa@unsa.edu.pe.

Declarations

Funding: This research received no internal or external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

Author Contributions

Miluska Odely Rodríguez-Saavedra: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing.

Robinson Bernardino Almanza Cabe: Methodology, Investigation, Data curation, Validation, Writing – original draft, Writing – review & editing.

Rubén Washington Arguedas Catasi: Software, Data curation, Validation, Supervision, Writing – original draft, Writing – review & editing.

References

[1] E. Bosio, S. Djankov, E. Glaeser, and A. Shleifer, “Public Procurement in Law and Practice,” American Economic Review, vol. 112, no. 4, pp. 1091–1117, 2022, doi: 10.1257/aer.20200738.

[2] M. Fazekas, I. J. Tóth, and L. P. King, “An Objective Corruption Risk Index Using Public Procurement Data,” European Journal on Criminal Policy and Research, vol. 22, no. 3, pp. 369–397, 2016, doi: 10.1007/s10610-016-9308-z.

[3] S. Williams-Elegbe, Fighting Corruption in Public Procurement: A Comparative Analysis of Disqualification or Debarment Measures. Oxford, U.K.: Hart Publishing, 2012, doi: 10.5040/9781472566232.

[4] T. Medina Arnáiz, “EU directives as anticorruption measures: Excluding corruption-convicted tenderers from public procurement contracts,” in International Handbook of Public Procurement, K. V. Thai, Ed. New York, NY, USA: Routledge, 2017, pp. 105–130, doi: 10.4324/9781315092539-6.

[5] G. Valverde, “El régimen de la caducidad en el ámbito sancionador del Organismo Supervisor de las Contrataciones del Estado: un análisis de la aplicabilidad de la caducidad en el procedimiento sancionador regulado en la Ley de Contrataciones del Estado del Perú,” Revista Eurolatinoamericana de Derecho Administrativo, vol. 10, no. 2, Art. no. e241, 2023, doi: 10.14409/redoeda.v10i2.13273.

[6] J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer, “Prediction Policy Problems,” American Economic Review, vol. 105, no. 5, pp. 491–495, 2015, doi: 10.1257/aer.p20151023.

[7] M. S. Lyra, B. Damásio, F. L. Pinheiro, and F. Bacao, “Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods,” Applied Network Science, vol. 7, no. 1, Art. no. 83, 2022, doi: 10.1007/s41109-022-00523-6.

[8] G. S. Becker, “Crime and Punishment: An Economic Approach,” Journal of Political Economy, vol. 76, no. 2, pp. 169–217, 1968, doi: 10.1086/259394.

[9] G. Spagnolo, “Reputation, competition, and entry in procurement,” International Journal of Industrial Organization, vol. 30, no. 3, pp. 291–296, 2012, doi: 10.1016/j.ijindorg.2012.01.001.

[10] J. V. Butler, E. Carbone, P. Conzo, and G. Spagnolo, “Past performance and entry in procurement: An experimental investigation,” Journal of Economic Behavior & Organization, vol. 173, pp. 179–195, 2020, doi: 10.1016/j.jebo.2020.02.019.

[11] M. P. Gerardino, S. Litschig, and D. Pomeranz, “Distortion by audit: Evidence from public procurement,” American Economic Journal: Applied Economics, vol. 16, no. 4, pp. 71–108, 2024, doi: 10.1257/app.20220512.

[12] D. Coviello, A. Guglielmo, and G. Spagnolo, “The Effect of Discretion on Procurement Performance,” Management Science, vol. 64, no. 2, pp. 715–738, 2018, doi: 10.1287/mnsc.2016.2628.

[13] J. Palguta and F. Pertold, “Manipulation of Procurement Contracts: Evidence from the Introduction of Discretionary Thresholds,” American Economic Journal: Economic Policy, vol. 9, no. 2, pp. 293–315, 2017, doi: 10.1257/pol.20150511.

[14] J. Gallego, G. Rivero, and J. Martínez, “Preventing rather than punishing: An early warning model of malfeasance in public procurement,” International Journal of Forecasting, vol. 37, no. 1, pp. 360–377, 2021, doi: 10.1016/j.ijforecast.2020.06.006.

[15] M. Fazekas and G. Kocsis, “Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data,” British Journal of Political Science, vol. 50, no. 1, pp. 155–164, 2020, doi: 10.1017/S0007123417000461.

[16] F. Decarolis and C. Giorgiantonio, “Corruption red flags in public procurement: new evidence from Italian calls for tenders,” EPJ Data Science, vol. 11, no. 1, Art. no. 16, 2022, doi: 10.1140/epjds/s13688-022-00325-x.

[17] M. Huber and D. Imhof, “Machine learning with screens for detecting bid-rigging cartels,” International Journal of Industrial Organization, vol. 65, pp. 277–301, 2019, doi: 10.1016/j.ijindorg.2019.04.002.

[18] M. J. García Rodríguez, V. Rodríguez-Montequín, P. Ballesteros-Pérez, P. E. Love, and R. Signor, “Collusion detection in public procurement auctions with machine learning algorithms,” Automation in Construction, vol. 133, Art. no. 104047, 2022, doi: 10.1016/j.autcon.2021.104047.

[19] S. L. Domingos, R. N. Carvalho, R. S. Carvalho, and G. N. Ramos, “Identifying IT purchases anomalies in the Brazilian government procurement system using deep learning,” in Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), 2016, pp. 722–727, doi: 10.1109/ICMLA.2016.0129.

[20] R. B. Velasco, I. Carpanese, R. Interian, O. C. G. Paulo Neto, and C. C. Ribeiro, “A decision support system for fraud detection in public procurement,” International Transactions in Operational Research, vol. 28, no. 1, pp. 27–47, 2021, doi: 10.1111/itor.12811.

[21] F. J. López-Iturriaga and I. P. Sanz, “Predicting public corruption with neural networks: An analysis of Spanish provinces,” Social Indicators Research, vol. 140, no. 3, pp. 975–998, 2018, doi: 10.1007/s11205-017-1802-2.

[22] E. Ash, S. Galletta, and T. Giommoni, “A machine learning approach to analyze and support anticorruption policy,” American Economic Journal: Economic Policy, vol. 17, no. 2, pp. 162–193, 2025, doi: 10.1257/pol.20210618.

[23] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[25] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," Information Fusion, vol. 81, pp. 84–90, 2022, doi: 10.1016/j.inffus.2021.11.011.

[26] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[27] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[28] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," Pattern Recognition, vol. 45, no. 1, pp. 521–530, 2012, doi: 10.1016/j.patcog.2011.06.019.

[29] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Trans. Knowl. Discov. Data, vol. 6, no. 4, pp. 1–21, 2012, doi: 10.1145/2382577.2382579.

[30] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: an analysis and review," International Journal of Forecasting, vol. 16, no. 4, pp. 437–450, 2000, doi: 10.1016/S0169-2070(00)00065-0.

[31] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. Mach. Learn., 2005, pp. 625–632, doi: 10.1145/1102351.1102430.

[32] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," PLOS ONE, vol. 10, no. 3, Art. no. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[33] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," Medical Decision Making, vol. 26, no. 6, pp. 565–574, 2006, doi: 10.1177/0272989X06295361.

[34] A. C. Cameron, J. B. Gelbach, and D. L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," Review of Economics and Statistics, vol. 90, no. 3, pp. 414–427, 2008, doi: 10.1162/rest.90.3.414.

[35] C. Molnar et al., "General pitfalls of model-agnostic interpretation methods for machine learning models," in xxAI – Beyond Explainable AI (Lecture Notes in Computer Science, vol. 13200), A. Holzinger et al., Eds. Cham, Switzerland: Springer, 2022, pp. 39–68, doi: 10.1007/978-3-031-04083-2_4.

[36] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," Journal of Clinical Epidemiology, vol. 110, pp. 12–22, 2019, doi: 10.1016/j.jclinepi.2019.02.004.

[37] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," ACM Computing Surveys, vol. 46, no. 4, pp. 1–37, 2014, doi: 10.1145/2523813.

[38] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: The Achilles heel of predictive analytics," BMC Med., vol. 17, no. 1, Art. no. 230, 2019, doi: 10.1186/s12916-019-1466-7.

[39] J. Gallego, M. Prem, and J. F. Vargas, "Predicting politicians' misconduct: Evidence from Colombia," Data & Policy, vol. 4, Art. no. e41, 2022, doi: 10.1017/dap.2022.35.

[40] A. Calveras, J.-J. Ganuza, and E. Hauk, "Wild Bids. Gambling for Resurrection in Procurement Contracts," Journal of Regulatory Economics, vol. 26, no. 1, pp. 41–68, 2004, doi: 10.1023/b:rege.0000028013.76488.44.

[41] F. Decarolis, “Awarding Price, Contract Performance, and Bids Screening: Evidence from Procurement Auctions,” American Economic Journal: Applied Economics, vol. 6, no. 1, pp. 108–132, 2014, doi: 10.1257/app.6.1.108.

[42] E. Colonnelli, S. Lagaras, J. Ponticelli, M. Prem, and M. Tsoutsoura, “Revealing corruption: Firm and worker level evidence from Brazil,” Journal of Financial Economics, vol. 143, no. 3, pp. 1097–1119, 2022, doi: 10.1016/j.jfineco.2021.12.013.

[43] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, “Human Decisions and Machine Predictions,” The Quarterly Journal of Economics, vol. 133, no. 1, pp. 237–293, 2018, doi: 10.1093/qje/qjx032.

[44] S. Kapoor and A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” Patterns, vol. 4, no. 9, Art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[45] M. Hino, E. Benami, and N. Brooks, “Machine learning for environmental monitoring,” Nature Sustainability, vol. 1, no. 10, pp. 583–588, 2018, doi: 10.1038/s41893-018-0142-9.

[46] M. O. Rodriguez Saavedra et al., “Machine Learning and Generative AI in Administrative Processes in Peru: Administrative Efficiency in the National Public Sector,” *Informatics*, vol. 13, no. 3, Art. no. 44, 2026, doi: 10.3390/informatics13030044.

[47] A. V. Morales Gonzales et al., “Normalization of corruption and evaluation of public services: adaptive mechanisms and subnational heterogeneity in Peru,” *Frontiers in Political Science*, vol. 8, Art. no. 1806986, 2026, doi: 10.3389/fpos.2026.1806986.