Scientific Reports

https://doi.org/10.1038/s41598-026-48873-w

Article in Press

Learning from sanctioned government suppliers: a machine learning and network science approach to detecting fraud and corruption in Mexico

Received: 22 December 2025

Accepted: 10 April 2026

Published online: 18 May 2026

Cite this article as: Medina-Hernández M., Kertész J. & Fazekas M. Learning from sanctioned government suppliers: a machine learning and network science approach to detecting fraud and corruption in Mexico. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-48873-w

Martí Medina-Hernández, János Kertész & Mihály Fazekas

Learning from sanctioned government suppliers: A machine learning and network science approach to detecting fraud and corruption in Mexico

Martí Medina-Hernández[1, *], János Kertész[1, +], and Mihály Fazekas[2, +]

[1]Department of Network and Data Science, Central European University, Vienna, Austria
[2]Department of Public Policy, Central European University, Vienna, Austria
[*]Corresponding author
[+]These authors contributed equally to this work

Abstract

Detecting fraud and corruption in public procurement remains a major challenge for governments worldwide. Most research to-date builds on domain-knowledge-based corruption risk indicators of individual contract-level features and some also analyses contracting network patterns. A critical barrier for supervised machine learning is the absence of confirmed non-corrupt (negative) examples, which makes conventional machine learning inappropriate for this task. Using publicly available data on federally funded procurement in Mexico and company sanction records, this study implements positive–unlabeled (PU) learning algorithms that integrate domain-knowledge-based red flags with network-derived features to identify likely corrupt and fraudulent contracts. The best-performing PU model on average captures 32% more known positives than random guessing and has a minimum of 20% precision on the top 5% of the predictions, substantially outperforming approaches based solely on traditional red flags. The analysis of the Shapley Additive Explanations reveals that network-derived features—particularly those associated with contracts in the network core or suppliers with high eigenvector centrality—are the most important. Traditional red flags further enhance model performance in line with expectations, albeit mainly for contracts awarded through competitive tenders. This methodology can support law enforcement in Mexico, and it can be adapted to other national contexts too.
Keywords: Fraud, corruption, measurement, public procurement, Mexico, positive - unlabeled learning

Introduction

The objective of this article is to develop a methodology that identifies and ranks contracts suspicious of fraud and corruption in public procurement (PP). This approach integrates three research strands: corruption risk indicators ("red flags") [13, 20], relational-based network information [14, 6], and positive-unlabeled (PU) learning [23] using real-world examples — i.e., government sanctions as indicators of fraud and corruption.

Public procurement refers to the process through which government agencies or state-owned enterprises purchase goods and services from private suppliers [33]. According to the OECD, PP represents about 12% to 29% of total government expenditure in its member countries [35, 34], an activity highly vulnerable to corruption [35]. Transparency International estimates that between 10% and 15% of PP spending is lost to corruption, though the real cost is likely higher [48]. Corruption in public procurement is particularly difficult to detect, monitor and prevent because the process is diverse, complex, information is highly asymmetric, and collusion between officials and bidders hides manipulation behind procedures that appear formally legal [34].

The development of corruption risk indicators or "red flags" capable of identifying suspicious contracts has been one of the main objectives in recent PP research. A major contribution in this direction was made by [13], where the Corruption Risk Index (CRI), a measure based on PP processes and contract characteristics was introduced. The CRI incorporates features known as **corruption risk indicators**

(or "red flags") that indicate corruption risks at the individual contract level, such as the presence of a single bidder, absence of a published tender call, a non-open procedure type, and a very short advertisement period, among others. Although the CRI effectively captures integrity risks embedded in specific contracts, it offers limited insight into how corruption emerges through the relationships among the actors involved.

To overcome this limitation, some researchers have increasingly applied **network science** tools to analyze corruption from a relational perspective [46, 29]. Typically, PP contracts are modeled as bipartite networks composed of two groups of nodes —government agencies (buyers) and companies (suppliers)— connected when a bid is submitted or a contract is awarded [12, 14, 46, 6, 31]. In other cases, the two groups represent companies and contracts [47], a structure used to detect cartel formation in bidding markets.

Although centralization and clustering in procurement networks are generally associated with corruption risk, their effects vary across countries and in magnitude. [46] found that corruption risk can concentrate either at the network's core or among peripheral actors, depending on the country. Similarly, [12] linked distinct clusters in PP networks to different levels of corruption and state capture, while [14] showed that high-risk agencies often have sparser network neighborhoods than expected. Other network measures, such as diameter, average path length, and average degree, have proven less informative for detecting corruption [24].

For both research approaches – domain-knowledge-based red flags and network science – establishing measurement validity and investigative usefulness have been a challenge due to the lack of proven cases of corruption as well as non-corruption, to validate against. In other domains, where learning data is more readily available, **supervised machine learning** techniques have been quite successful, such as online auction sites [2] or shill bidding detection (when participants place fake bids to artificially raise prices) [17]. In the few cases when labeled data has been used in PP, two major challenges remain. First, the labels employed are often not actual instances of corruption but rather likely cases or grey cases [15] or sometimes even using proxy indicators to flag cases . One solution to this problem could be the use of debarments, sanctions or administrative penalties issued by national authorities as positive labels in machine learning models. Although such information is publicly available in some countries [38], its use has been a largely underexplored in PP research. Importantly, public authorities issuing sanctions may themselves be corrupt or under political control so the derived labels may be biased requiring a careful assessment of data quality. Moreover, sanctions are all but one of many types of hard evidence of fraud and corruption with further examples including corruption convictions against company owners [7]. Using company sanctions as PP labels in machine learning models face three practical challenges: (1) mapping company-level sanctions to contract-level labels, (2) the absence of negative observations, and (3) data imbalance. Typically, as in this study, data on sanctioned companies do not include the specific contracts involved but only general information about the sanction. Therefore, it is necessary to define criteria for assigning labels to individual contracts, affecting both the training sample and its representativeness. Companies with many contracts will receive more labels, biasing model outcomes. Moreover, using sanctions as positive labels implies the lack of true negative examples, violating the assumptions of traditional Positive-Negative (PN) learning and distorting standard performance metrics. In addition, because only a small fraction of firms are sanctioned, any PP dataset will contain very few positive cases.

The second challenge for applying supervised machine learning to fraud and corruption detection is methodological. The work of Aldana et al. [1] represents an important advance in this respect. However, the aforementioned challenges are not fully met. Although labeling all contracts from a sanctioned company as corrupt is plausible, it can overrepresent large companies and risks data leakage without a stratified company-level train-test split. Furthermore, the assumption that all non-labeled companies are non-corrupt treats the problem as PN learning, when in reality only positive labels exist, while in reality the problem represents an example of the **Positive-Unlabeled (PU) learning** challenge and should be treated accordingly. The same applies to PP machine learning research using criminal convictions [7] or political connections [45]. Like traditional classification, PU learning aims to distinguish positive from negative cases, but without explicit negative labels [23]. This limitation is non-trivial, as the predictive performance of PU classifiers strongly depends on label availability [28, 40]. There is no standard method for PU learning, though several strategies have been proposed; comprehensive reviews can be found in [23, 22]. Most have been tested on balanced datasets, while only a few target imbalanced ones [36]. We focus on two algorithms specifically designed for this scenario: the Positive-Unlabeled Bagging (PU Bagging) [30] and the Hellinger Distance Stratified Random Forest (HDSRF) [36].

PU Bagging assigns misclassifications costs by aggregating classifiers trained to distinguish positive from unlabeled samples and averaging their predictions. Training uses bagging, a resampling strategy

that draws bootstrap samples from the unlabeled data while including all labeled positives. This method handles imbalanced datasets naturally and can be combined with various classifiers, most often Support Vector Machines (SVMs) [30, 36].

A simpler yet effective alternative is the **HDSRF** algorithm, which simultaneously addresses PU learning and class imbalance. It employs the Hellinger distance as the split criterion in the random forest's decision trees, assuming a positive class prior larger than the observed positive proportion. In addition to bagging and random feature selection, the algorithm ensures that all positive samples appear in each bootstrap iteration (hence, "stratified"). The Hellinger distance has been shown to outperform Gini and Entropy for imbalanced datasets, though it requires knowing the number of positive samples —a limitation that can be mitigated by treating the class prior as a tunable hyperparameter [36].

Both HDSRF and PU Bagging assume that positive examples are Selected Completely at Random (SCAR), meaning the observed positives are a random subset of all positives in an unobserved PN dataset. This assumption is rarely valid and remains one of the key challenges in PU learning [23] and it is also a problem in our case.

Earlier work has made substantial progress in identifying corruption risks using a diversity of approaches. We advance this literature in a number of ways: First, we show that widely available data on sanctioned suppliers can serve as labels to infer risk factors from. Second, we implement a robust and replicable method, tailored to the typical prediction task expected in other countries' PP datasets. This recognizes the PU structure and imbalanced nature of the data. We also introduced a method for evaluating PU-learning tasks that is useful for ranking instances and robust to cases where predicted probabilities are similar. Third, this paper combines, confirms and refines red flags from prior literature such as direct awards; while it also incorporates a wide array of network features derived from the contracting network such as eigenvector centrality. Importantly, established red flags are most precise when aggregated to the supplier level, rather than on the contract-level. Finally, our results also point at the practical usefulness of tailored machine learning models for detecting fraud and corruption in practically relevant real-life settings.

Data and Methodology

Datasets

The **contracts dataset** includes 2,301,278 contracts awarded by Mexican federal, state, and municipal governments between 2011 and 2022 using federal funds. It contains contract-level information for 5,304 government entities and 259,534 private suppliers. The dataset was compiled from contract- and procedure-level data available on the Mexican government's CompraNet platform [4], following [11, 8]. As of the time of writing, the government's original "CompraNet" website is no longer in operation and has been superseded by the "Compras MX" platform. Full details on the original datasets and the matching process are provided in Supplementary Information A.

To generate labels for our model, we created the **sanctions dataset**, comprising 14,535 unique companies. This dataset combines two sources: 12,396 companies identified for issuing invoices for simulated operations ("Empresa que Factura Operaciones Simuladas" or EFOS) [42] and 2,139 companies sanctioned for involvement in corrupt activities in public contracting ("Proveedores y Constratistas Sancionados" or PCS) [41]. Of these, 1,673 companies had contracts in the contracts dataset, with 748 from EFOS and 925 from PCS. More details on the data sources and datasets can be found in Supplementary Information A. Crucially for the trustworthiness and reliability of our labels, companies sanctioned in the EFOS and PCS datasets are determined by the Tax Administration Service ("Servicio de Administración Tributaria", in Spanish) and the Ministry of Public Administration ("Secretaría de la Función Pública") respectively, formally independent from both the contracting authorities and suppliers in the dataset. We consider potential political bias in the labels as relatively contained, because our data spans across two different federal governments, hence a new administration could investigate wrongdoing by the predecessor. Nevertheless, administrative biases may remain, due to the different auditability or the difficulty of investigation of different procurement procedures. This is a point we will return to when discussing our detailed results.

We denoted all contracts of sanctioned companies as fraudulent, regardless of when the sanction was made or the contract signed. Including all contracts of all sanctioned suppliers results in the most comprehensive learning dataset possible, encompassing a wide variety of corrupt behavior. Nevertheless, this is a strong assumption and it implies that sanctions did not have an effect on subsequent company behavior and that contracts of the same company show strong similarity across time. We tested this

assumption by reviewing the sanctions' impact on company behavior, checking the rate at which suppliers are sanctioned more than once and comparing our chosen model with models based on alternative labeling assumptions (for full details see Supplementary Information B. Specifically, we found little to no change in the prevalence red flags used in the literature (e.g. non-open procedure types) for sanctioned companies from before to after the sanctions. This also holds when comparing all distinct years of sanctioned supplies with each other. Moreover, a non-negligible share, 13% of suppliers which receive contracts after being sanctioned (17%) get sanctioned again. Finally, models based on labeling sanctioned suppliers' contracts differently after they are sanctioned are compared to our preferred model. This reveals that our model is generally more robust and performs relatively better, or at least as well as the others.

Figure 1 shows the yearly distribution of sanctioned contracts after matching with the contracts dataset. After an initial rise in contract numbers, there is a clear decline from 2018, coinciding with the end of the administration of Enrique Peña Nieto (EPN) and the early years of Andres Manuel López Obrador (AMLO). Contract numbers rise again in 2021. The proportion of contracts considered fraudulent (our positive labels) remains between 2.2% and 5%.

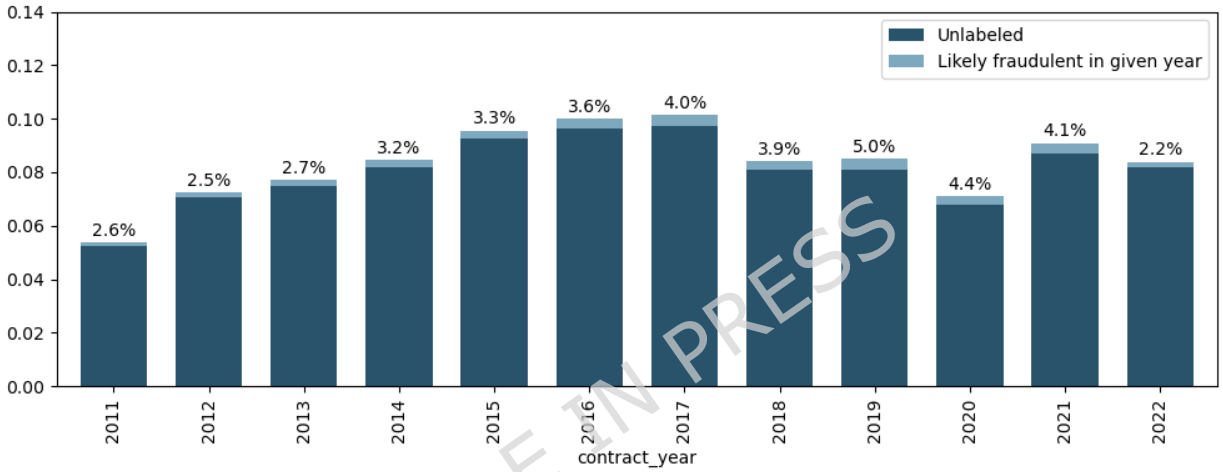

Figure 1: Distribution of contracts and labels across years. The height of each bar represents the percentage of contracts (y-axis) in a given year (x-axis) relative to the total number of contracts in the dataset. The dark blue in the bars represent the unlabeled contracts in the dataset, meanwhile the light blue rectangles and the percentages at the top of the distribution indicate the percentage of positive labels in that specific year.

Feature engineering

In line with our research purpose, we included two types of features in the model: domain-knowledge features, which capture well-known indicators of fraud and corruption, and network features, derived directly from the bipartite network and its transformations.

Domain-knowledge features are defined at the contract, buyer, and supplier levels and assigned to each contract in the dataset. Some are directly extracted from the contracts dataset, including contract price, legal framework, supplier size, and supply type. Most features are based on established proxy indicators of corruption, such as single bidding, non-open procedure type, short submission period, short decision period, and Benford's Law which are also combined into a simple composite score: the Corruption Risk Index (CRI) [13].

The CRI captures the corruption risks of public contracts by combining different theoretically motivated indicators of corruption available in the dataset. These indicators are validated through statistical testing based on their relationship with one of the most reliable observable signals of corruption: single bidding, defined as a tender in which only one bidder participates. This statistical testing is performed using a logistic regression model, where the dependent variable is single bidding and the independent variables are other indicators of corruption; furthermore, contract year, contract price decile, supply type, and region are used as controls. Indicators that emerge as significant predictors are the statisti-

cally validated corruption risk indicators, or red flags, as they are likely to indicate deliberate restriction of competition in otherwise competitive public tenders. By implication, red flags are identified and parametrized independently of our machine learning models. The regression coefficients are used to assign equally separated weights (such as 0, 0.5, 1) to the red flags, that are subsequently averaged to create the CRI. The exact procedure for defining individual red flags and calculating the composite CRI is detailed in Supplementary Information E. From a modeling perspective, note that while individual risk indicators such as single bidding are entered in the model on the contract and aggregate levels, the CRI is only included in aggregated form (e.g. at the supplier, buyer, or edge level), not as a contract-level variable. This set-up avoids redundancy between the composite CRI and individual red flags, allowing the model to learn from measures of corruption beyond the contract level.

We also included corruption indicators adapted for the Mexican context, inspired by the work of [20], such as suppliers not being recorded in the National Registry and Fragmented Contract Odds[20]. Additionally, we included the variables Buyer-supplier active weeks, Buyer-supplier number of contracts per week, Buyer-supplier spending per active Week, which were previously used to analyze corruption in Mexico [8]. Based on the contract-level risk indicators, we also computed aggregated measures, including the average CRI of supplier-buyer edges and CRI of an edge's neighborhood. Taken together, this set of proxy indicators of corruption in public procurement are comprehensive and offer a detailed picture, even though some key indicators remain imperfect, such as single bidding which suffers from over 85% missing rate.

Another set of features are derived from network science. Buyer-supplier relationships were represented as a bipartite network, with links between actors connected by a contract and weights equal to the number of such contracts. In total, we generated one bipartite network per year, in order to represent suppliers' and buyers' positions over time. Variables such as Coreness [46], Edge Betweenness Centrality, and Competitive Clustering [14] were calculated for every year.

We also derived features from the buyer and supplier projections of the bipartite network for every year, including Betweenness, Closeness, Degree, and Eigenvector Centrality. All network centralities were calculated using normalized formulations in the `igraph` Python package [5].

All network and domain-knowledge features were determined on an annual basis, such that each value corresponds to a given year. Supplementary Information C provides definitions for all 60 features, including domain-knowledge and network-derived features from three network types: the bipartite network and its supplier and buyer projections. Additional characteristics, such as general statistics, distributions, and missing values, are reported in Supplementary Information D.

As described in Supplementary Information D, 12 out of 60 features present some degree of missingness. The missing values of some of them, such as Compliant Submission Period, Tender Period, Legal Fundament, are related with the type of procedure: the former two are only defined for non-direct procedures, meanwhile the later is defined only for direct ones. MAD (Mean Absolute Deviation to Benford's distribution) is calculated only for buyers with at least the equivalent of the quantile 0.75 number of contracts for all buyers in a given year, which is equivalent to a minimum of 42 - 62 contracts, depending on the specific year. Therefore, the contracts with missing MAD are those that belong to buyers with small number of contracts. The missing values of Neighborhood Avg. CRI and Neighborhood Prop. Recorded-Direct Procedures correspond to links with no neighborhoods, i.e. disconnected components in the bipartite buyer-supplier network. Therefore, the only variables with missing values not created by design, are Legal Framework, Procedure venue, R.F. Procedure Type, R.F. Single Bidder, Supplier Size and Supply Type. Since categorical variables were transformed using one-hot-encoding, we also included a "missing" variable for each one of them. In the case of continuous variables, we assigned out-of-range values to indicate missingness for each feature, a common practice in feature engineering [16].

Train-test split

Given the characteristics of our dataset, a simple random-sampling train-test split was not feasible. Our analysis shows that 5% of suppliers account for nearly 64% of the contracts, with a Gini coefficient of 0.77 (see Supplementary Information Figure F). This violates the independent and identically distributed (IID) assumption of many machine learning models, including those relying on bootstrap sampling, such as HDSRF and PU Bagging. These methods draw subsamples of unlabeled observations, which are likely biased toward larger companies, underrepresenting smaller ones. Furthermore, contracts from the same supplier exhibit high similarity (see Supplementary Information F-Figure 7), and all contracts of a sanctioned supplier are labeled positive. Consequently, models overrepresenting companies with many contracts would learn to identify large suppliers rather than general patterns of fraud.

To address these issues, we implemented a company-based, undersampled train-test split. Companies in the training set do not appear in the test set, and companies with many contracts are undersampled in training. Details are provided in Supplementary Information F. On the contrary of other machine learning papers that deal with imbalanced data, we did not balance our test set in order to give a realistic performance of our models in real scenarios, therefore the test set remained untouched.

We explored two learning approaches to investigate the usefulness of our method: the transductive and inductive learning. The transductive learning objective is to detect observations from an already established sample. The inductive learning objective is to detect observations of data that has not been seen. The transductive learning setting would help authorities to detect past contracts, while the inductive setting would be useful for future ones.

Our main objective is framed in transductive learning, aimed at detecting fraud within the current contracts dataset, so the primary train-test split does not consider temporal aspects of the policy of treating fraud in PP, even though the sanctions dataset spans two administrations —EPN and AMLO—which may have applied different strategies.

To evaluate this, we also tested our models on two administration-specific datasets aiming to detect future fraudulent contracts (inductive learning). For each, we trained on the early years of each administration and predicted the second-to-last year of contracts (2017 for EPN, 2021 for AMLO), hiding sanctions from future years in the training set to simulate a real-world scenario where authorities only have past sanctions.

PU Learning Models and validation

Since our dataset is imbalanced and contains positive and unlabeled instances, we used two algorithms suited for the this task: Hellinger Distance Stratified Random Forest (HDSRF) [36] and PU Bagging with an SVM learner [30]. Both algorithms specifically deal with imbalanced labels [32].

Experiments were conducted in Python using the scikit-learn framework [37]. Given the relatively recent adoption of PU learning methods, we relied heavily on the publicly available code from [36], which provides implementations of several PU learning algorithms, including those used in our study.

We performed hyperparameter tuning and evaluated performance with 4-fold cross-validation. Each fold was stratified using the company-based, under-sampled train-test split, ensuring no company appeared in both training and test sets. A fifth set of contracts, not included in the cross-validation folds, was reserved for probability calibration.

One concern of the use of the HDSRF is the assumption of knowledge of class prior. In our case, we used it as a tunable parameter since there is no reliable information on its approximate value. However, we tested different class priors and concluded that the results are largely insensitive to the specific value of the prior (see Supplementary Information G).

Performance evaluation and model selection

One of the main challenges in PU learning is model evaluation. In traditional imbalanced supervised binary classification, evaluating the performance of a model is based on traditional confusion-matrix-based metrics, such as *Recall*, *Precision* and $F1$-*score*, defined as:

$$Recall = \frac{TP}{FN + TP}; \quad Precision = \frac{TP}{TP + FP}; \quad F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{1}$$

where TP are true positives, FN are false negatives, and FP are false positives, defined over a probability threshold. In PU learning, however, none of these quantities are certain since we do not have reliable negative labels and their interpretation should be done with caution. For instance, some authors have argued that *Recall* in a PU setting could be considered as approximation in a PU setting if the observed positive labels are follow the SCAR assumption [21]. Moreover, *Precision* in a PU setting should always be considered as an underestimation of the real metric, since any FP observation could actually be an unlabeled positive. The uncertainty of *Recall* and *Precision* translates directly to $F1$-*score* and average precision score, since both are based on trade-offs between both measures.

Another machine learning area where there is a presence of unlabeled observations is recommender systems, where the task is to identify potentially relevant items for a user that has not observed them. Even when in principle, training and testing recommender systems include labeled negatives, a common way to understand the performance of the model is by assuming that a top percentile of items rated by a user is relevant and use traditional metrics such as *Recall* and *Precision* [19]. This approach has

two caveats: First, it is dependent on the proportion of positives in the sample which is unknown when deciding about the top percentage to be taken. Second, the division of data is based on an unambiguous order of observations, i.e., it assumes a perfect ranking of instances.

The recommender systems evaluation approach aligns with our purposes, given that the ultimate objective of our model is to identify a small subset of contracts for authorities to review, moreover, we can also focus on the ranking quality of known positive contracts. However, we need to deal with the dependence of proportion of positives and the assumption of perfect ranking of instances. As a methodological contribution, we propose to evaluate the model through a tie-aware cumulative sum function, that can be translated to robust versions of recall and precision in the top % of instances. Additionally, since we are interested in the ranking properties of the model, we propose to compare the models with a null model based on random guessing that would help us measure how different is our model from random guessing with perfect ranking. For this, we also adapt the concepts of gain and lift [3].

Let $\{(y_i, \hat{p}_i)\}_{i=1}^n$ be the set of n true labels $y_i \in \{0, 1\}$ and predicted probabilities $\hat{p}_i \in [0, 1]$, ordered such that $\hat{p}_1 \geq \hat{p}_2 \geq \cdots \geq \hat{p}_n$. Let $P = \sum_{i=1}^n y_i$ be the total number of known positives, and $\pi = P/n$ the prevalence.

The cumulative number of positives up to instance k is:

$$C(k) = \sum_{i=1}^{k} y_i.$$

The biased recall and precision at instance k are respectively defined as:

$$R@k = \frac{C(k)}{P}; \quad P@k = \frac{C(k)}{k}, \tag{2}$$

assuming that the classifier has assigned a positive label to all ordered instances up to instance k. If different instances have equal probabilities there is no unbiased ranking and therefore $C(k)$ cannot give us a strict metric. Let $\mathcal{G} = \{g_1, g_2, \ldots, g_m\}$ denote the partition of indices $\{1, \ldots, n\}$ into groups of instances with equal predicted probability. For each $g_j = [a_j, b_j]$, we have $\hat{p}_{a_j} = \hat{p}_{a_j+1} = \cdots = \hat{p}_{b_j}$ and $\hat{p}_{b_j} > \hat{p}_{b_j+1}$.

The *robust tie-aware cumulative sum* $C_R(k)$ is the instance-based cumulative step function:

$$C_R(k) = \begin{cases} C(a_j) & \text{if } a_j \leq k < b_j \\ C(b_j) & \text{if } k = b_j \end{cases} \quad \text{for } k = 1, \ldots, n. \tag{3}$$

We can define then robust recall at instance k as:

$$R_R@k = \frac{C_R(k)}{P}; \tag{4}$$

and precision at instance k as:

$$P_R@k = \frac{C_R(k)}{k}; \tag{5}$$

The robust recall $\frac{C_R(k)}{P}$ is conceptually similar to a "recall curve", which plots recall (True Positive Rate) at every classification threshold based on unique predicted probabilities. The key difference is that $C_R(k)$ is defined by instances, not thresholds, and only after all instances of similar probabilities are accounted for. This allows us to distinguish between a classifier with a smooth probability distribution and one that assigns blocks of equal probabilities, which is particularly useful for auditing purposes. Moreover slicing the predicted instances at an instance k using a biased recall or precision, would be misleading if there are instances with similar probability, since it would count as positive all instances with same predicted probability, even if the true values are not positive. This approach also prevents the common issue of assigning a recall value of 1 to classifiers that predict all instances as positive. In our modified gain, all instances except the last receive zero gain; therefore, the average gain for a classifier that predicts every instance as positive will be close to zero.

Given the limitation of these metrics for PU learning, and that we are interested in how good the classifier is to capture known positives, we decided to compare it with a *null model*. The null model for the previous functions, assuming a perfect ranking (each instance has a different probability), is the expected cumulative positives under random guessing:

$$C_0(k) = \pi k.$$

Following [3], we compare our tie-aware cumulative function with the null model by using gain and lift. The *normalized gain* at instance k is:

$$\text{Gain}(k) = \frac{C_R(k) - C_0(k)}{P},$$

and the *average gain* over all instances is:

$$\overline{\text{Gain}} = \frac{1}{n}\sum_{k=1}^{n} \text{Gain}(k). \tag{6}$$

The lift [3] measures how much better the predictions are compared to a baseline, and is defined as the ratio of the probability that the instance is a positive given that the model predicted that it is positive (in our case, $\frac{C_R(k)}{k}$), and the probability that the null model predicts it as positive ($\frac{C_0(k)}{k}$) . Therefore, the *lift curve* at instance k is:

$$L_c(k) = \frac{C_R(k)}{C_0(k)} \quad \text{for } k = 1, \ldots, n. \tag{7}$$

The *average lift* is

$$\overline{L_c} = \frac{1}{n}\sum_{k=1}^{n} L_c(k), \tag{8}$$

representing the average factor by which the classifier captures more known positives than random guessing.

Our robust tie-aware cumulative sum $C_R(k)$ is defined as a step function over instances of equal predicted probabilities. The gain at a specific instance is the difference between the classifier's $C_R(k)$ and the null model $C_0(k)$, while the lift is the ratio between $C_R(k)$ and $C_0(k)$. Average gain measures the proportion of known positives captured beyond random guessing, and average lift measures the factor of improvement over random guessing. Unlike standard average gain and average lift [3], our method calculates gain only after all instances with the same probability are accounted for, making it robust to classifiers that assign identical probabilities to multiple instances.

Our selected model is also tested for significance using permutation testing [3] indicating the probability of observing results as extreme as our model, given that the null hypothesis is true.

Interpretation of the model

While the primary objective of this article is to introduce a methodology that identifies and ranks contracts suspected of fraud and corruption, we also examine the model in detail. This is done to demonstrate the analytical value of our approach without engaging in formal theory testing. Given the law enforcement context in which the model is applied, interpreting the features used during training is important not only for ensuring transparency but also as guidance for auditing authorities. The interpretation of the model was done through SHapley Additive exPlanations (SHAP values) for tree-based models [25], an efficient method for the exact computation of Shapley values [43]. Shapley values provide a principled way to locally attribute each considered feature's contribution to an individual machine learning prediction [27, 26]. Specifically, we focus on the average SHAP values of features, which serve as a measure of feature importance for the model. Additionally, we analyzed the distribution of SHAP values for individual features to assess the direction of their influence –that is, whether they contribute positively or negatively to the model's predictions.

Results

Performance

The performance of the models can be observed in figure 2. In the top row of the panel we observe the robust recall ($R_R(k)$), in the middle row the robust precision ($P_R(k)$), while in the bottom row the robust lift ($L_R(K)$ is shown. The blue curves in the figures correspond the average performance of the model at instance k over 4 cross validation subsets of the data. The first two columns of this panel correspond to the PU Bagging and HDSRF models in the transductive setting. Since extensive research has been done in the creation of the "red flags", a natural comparison with our models is to rank the instances with CRI [13] and evaluate how different it is from our machine learning models. The comparison can be observed in the most right column.

The figure 2 clearly shows that the HDSRF outperforms PU Bagging as well as ranking by CRI in both the robust recall, robust precision and the lift. By taking the top 5% of instances, the HDSRF captures 30% of the known positives, and has a robust precision of 20%, which represents 25% more known positives (gain) and 6 times more than than random guessing. Taking into account the whole distribution of instances, on average, the HDSRF captures 32% more known positive instances (average gain), and 2.31 times more known positives (average lift) than random guessing. In comparison, the PUBagging underperforms in all these metrics, with a robust recall, precision and lift of zero up to the top 55% instance. See Supplementary Information H.1 for the performance results over all distribution of instances.

The underperformance of PU Bagging can be explained by its tendency to assign the same probability to many test instances. The sudden rise in gain around 50% of instances occurs because gain is calculated over groups of instances with equal probability; PU Bagging assigned the highest predicted probability to that many instances (see H.2 in Supplementary Information). This behavior suggests that PU Bagging tends to over classify instances as positive, limiting its usefulness for ranking. In contrast, HDSRF exhibits a smoother, steadily increasing performance.

For auditing purposes, these performances mean that if authorities audit the top 5% of the contracts in the dataset according to the predicted probabilities of the HDSRF, at least 20% of them will be fraudulent or corrupt ($P_R@k$). It is important to highlight that this figure is always an underestimation of the true performance of the model, since the instances categorized as false positives could be unlabeled positive cases. This is why it is important to compare it with a null model, in our case, random guessing to asses how much better is at capturing known positives in comparison with random guessing in the whole distribution.

Moreover, our proposed evaluation method also can point out the usefulness for ranking. For instance in figure H.3 from Supplementary Information we can observe the biased (not corrected for instances with same probability) versions of recall, precision and lift at instance k, where PU Bagging performance is much better, but hiding its tendency to classify many instances with same predicted probability, hurting its usefulness for the auditing authorities.

One might argue that contract concentration in the test set could inflate performance metrics. To address this, we also evaluated performance on a uniform test set, following the undersampling procedure described before. As shown in Supplementary Information H.4, the performance is very similar, indicating that it is independent of contract concentration in the test set.

We calculated the same metrics also for the inductive setting, showing very similar results: for the EPN administration, we got an average gain of -0.05 and 0.31 and an average lift of 0.62 and 2.31 for PU Bagging and HDSRF, respectively. For the AMLO administration, average gain of -0.2 and 0.29 and average lift of and 0.36 and 1.81, also for PU Bagging and HDSRF respectively (see Supplementary Information H.5). In both administrations, the HDSRF is significantly better than the PU Bagging, and the performance is very similar to the transductive setting.

Given that the HDSRF showed the best results for average gain and average lift, we proceeded to test for significance through random permutation tests [3]. The results indicate that the results of our model are out of the distribution of the permuted labels for average gain and average lift (see Supplementary Information H.6).

Feature Importance

Given the significantly better performance of HDSRF, we focus on this model for feature analysis and interpretation. The mean SHAP values of HDSRF are shown in Figure 3. Panel (a) presents the top

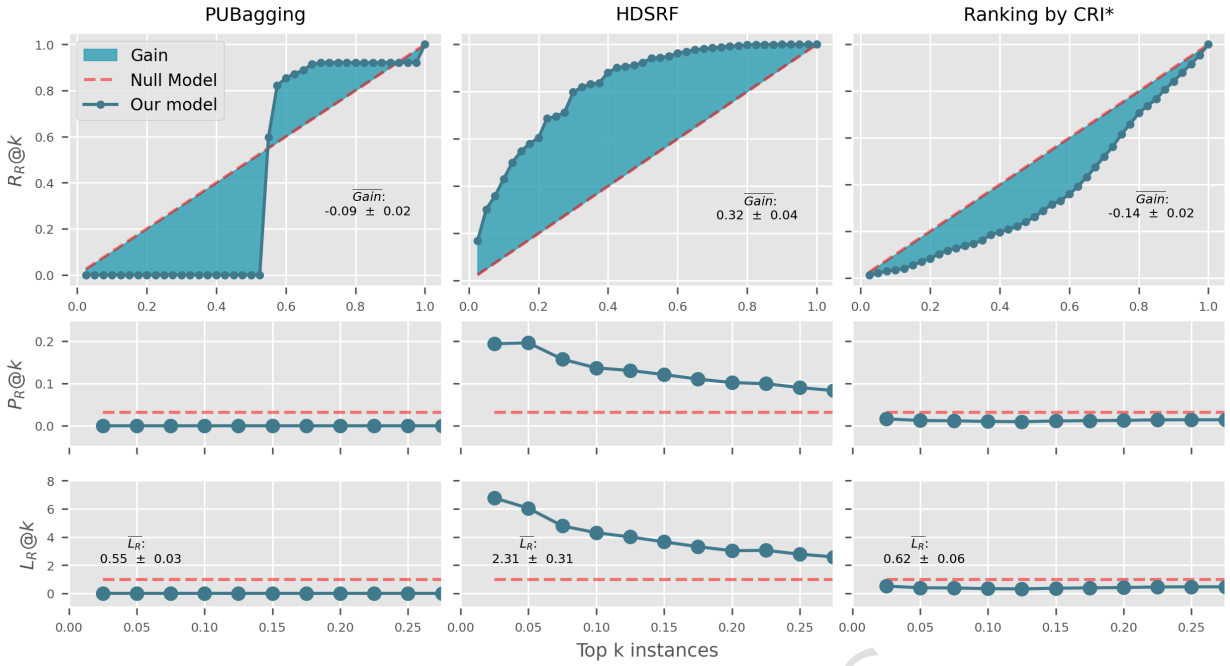

Figure 2: Model's performance (robust metrics). Each blue line corresponds to the average robust recall ($R_R@k$), robust precision ($P_R@k$) and robust lift (($L_R@k$)) at instance k over the 4-fold cross-validation sets for the examined models in comparison with a random classifier (red dashed line). The Colored area corresponds to the gain: the difference between the $R_R@k$ and the null model. The figures indicate that the HDSRF outperforms the PU Bagging model for every set. Moreover, the straight line at value zero up to $\sim$ 0.5 normalized ranks in the PU Bagging algorithms shows that around 50% of the observations in the test set are classified with the highest predicted probability, which makes unfeasible to work for ranking purposes. *Ranking by CRI implies to rank the contracts using only the CRI and estimate how well it captures labeled contracts without any machine learning methodology.

most important features according to their absolute mean SHAP value. After feature engineering, the model includes 69 features, 9 of them resulting from one-hot encoding. Supplier Coreness (weighted degree), Supplier Eigenvector Centrality, and Supplier Proportion of Recorded Direct Procedures are the top three features, highlighting the importance of a supplier's network position and contract composition in detecting fraud.

Panel (b) shows the grouped sum of absolute average SHAP values by feature type. Grouped, network features contribute most to the predicted probability of a contract being corrupt. Although domain-knowledge features are rarely among the very top, taken together they still have a substantial impact on the predicted probability. Among the top 30 features, domain-knowledge indicators typically feature as company-level aggregates of contract-level measures, such as the proportion of recorded direct and post-direct procedures and average CRI.

Panel (c) provides an overview of feature type rankings across all variables. The first 10 features are a mixture of all types, while most of domain-knowledge dimensions appear at the lower end of the mean SHAP value ranking.

We also compared transductive and inductive learning settings across the EPN and AMLO administrations. Figure 4 shows non-absolute SHAP values for each feature per instance, colored by feature values. All plots display the average SHAP values of 30 features, ordered by importance in the transductive setting. Empty spaces indicate that a feature is not present in the top 30 for that setting.

Several observations arise from this comparison. We see general stability among the top 30 features across the three settings. First, the three settings share 22 features out of 30 and the direction of their relationships is largely consistent. Additionally, all features present in the inductive setting are shared across both government administrations. Although the mean SHAP values are generally smaller in the inductive setting compared to the transductive one, the relative impact of each feature is similar. It

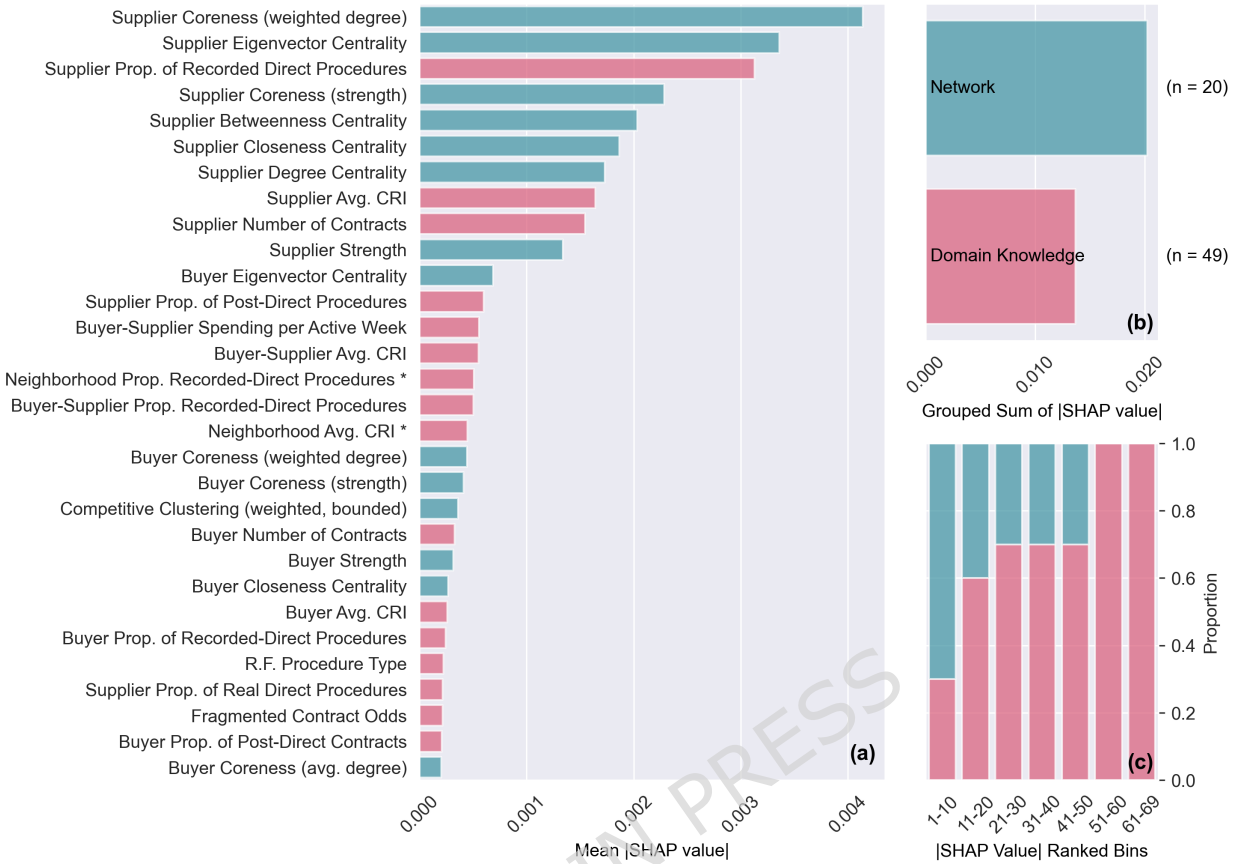

Figure 3: Top 30 most important features of the model. In (a) we observe the top 30 most important features of our HDSRF according to their mean absolute SHAP value, where the top 10 features are dominated by network features (7 out of 10). The accumulated influence of different variable types is shown in (b). The network features of the model have the highest sum of absolute mean SHAP values. The position of individual features over the entire set of features is depicted in (c). The top positions are a mixture of network and domain-knowledge features, meanwhile the bottom ones are dominated by domain-knowledge features, even though they are present in all of the ranked bins. The definition of each feature can be read in Supplementary Information C. * These are features that are take domain-knowledge information by leveraging the network structure

is worth noting that some highly correlated variables, such as those related to Competitive Clustering, appear in the inductive setting but not in the transductive one, or appear in the transductive but not in the inductive. This suggests that a method based on SHAP values may have difficulty distinguishing the contributions of highly correlated features.

In addition to the SHAP values, we also generated an ablation study that train and test the performance of the HDSRF for the two types of features: domain-knowledge features and network features. The performance of the models, specially at the top %k instances, favour the network features more than the domain-knowledge. See Supplementary Information I.1 for more detailed analysis.

Model Interpretation

The most relevant model features were identified based on their absolute mean SHAP values and theoretical importance. Their corresponding SHAP dependence plots were subsequently analyzed (Figures 5 and 6). These plots convey four types of information. First, they show the relationship between a feature's value and its contribution to the predicted probability of an individual instance (SHAP value). Second, the Colors indicate their association with selected other features used in the model. Vertical

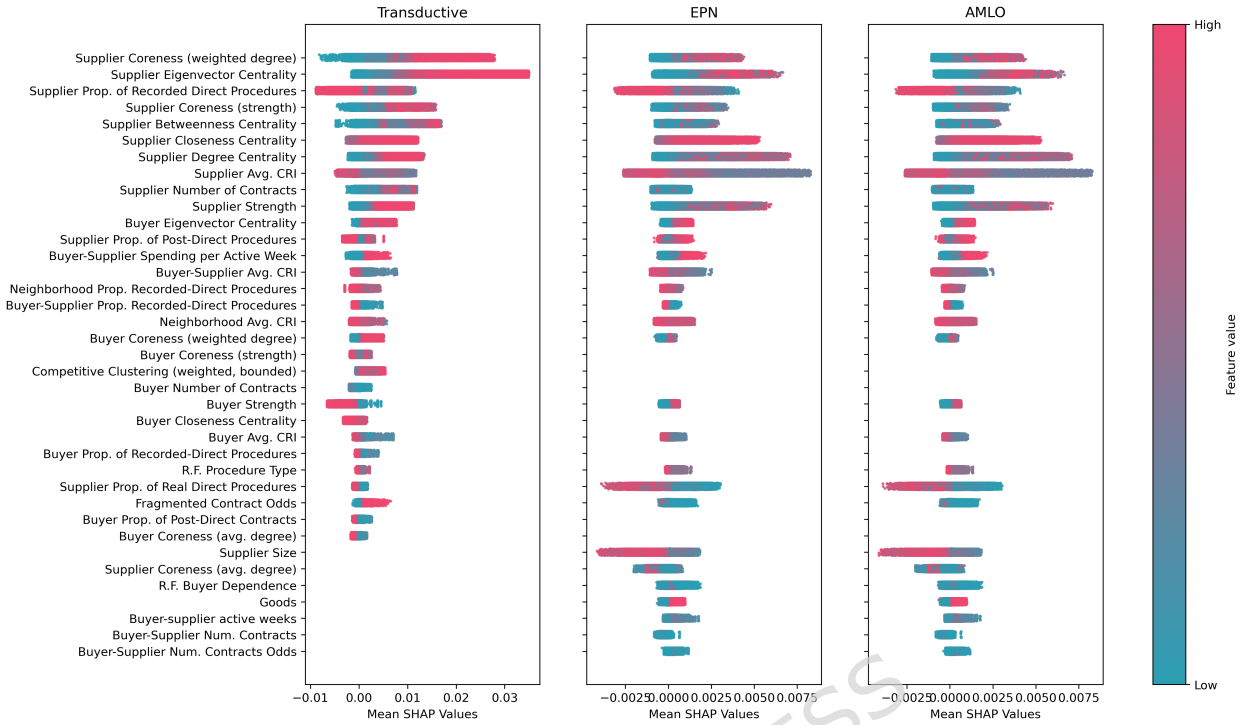

Figure 4: Mean SHAP values of top 30 across transductive and inductive (EPN and AMLO administration) learning. We can observe the non-absolute SHAP values of each feature per instance colored by the values of the feature. All of them shows the average SHAP values of 30 features ordered by the absolute SHAP value in the transductive setting. The empty spaces show that a feature is not present in the top 30 for that setting.

Color changes suggest feature interactions, while horizontal Color changes indicate correlations. With interaction, we mean that the effect of feature A on the dependent variable (expressed as SHAP values) changes depending on the values of feature B. Third, the grey band centered around zero in the SHAP values denotes the boundaries corresponding to the 10th and 90th quantiles of the SHAP value distribution across all features. This provides an intuitive indication of how common or uncommon specific SHAP values are. These boundaries do not represent statistical significance but rather indicate the relative extremity of SHAP values within the overall distribution. Fourth, the blue histogram in the background illustrates the distribution of the feature values.

Network features

In Figure 5, we show the SHAP dependence plots of ten network features from the model and the Pearson correlation (ρ) between the SHAP values and the original values of the model. These features capture the positions of suppliers and buyers in the bipartite network (for coreness) and in their respective projections of the Mexican public procurement network. Supplier SHAP values are colored by the corresponding buyer features, and vice versa.

Both supplier and buyer **coreness** show a positive correlation with the SHAP values, meaning that contracts involving actors closer to the bipartite network core receive higher positive contributions to the probability of being fraudulent. For suppliers, this relationship is consistent regardless of the buyer's position, while for buyers, SHAP values are only high when the contracts are signed with suppliers near the core. Supporting this observation, supplier and buyer **eigenvector centralities** also display a positive correlation with SHAP values. In the supplier projection, high supplier eigenvector centrality indicates shared buyers with high-degree suppliers, forming a densely connected supplier core. Similarly, high buyer eigenvector centrality reflects shared suppliers with well-connected buyers. When buyer eigenvector centrality is low but supplier eigenvector centrality is high, the contribution becomes slightly

negative, though with small magnitude.

The similar positive correlation of coreness and eigenvector centrality with their SHAP values, together with a Pearson correlation of 0.68 between supplier coreness and supplier eigenvector centrality, suggests a clear core structure in the network where fraudulent contracts are concentrated. Such a concentration of high risk actors in the center of the contracting network is consistent with the phenomenon of state capture [12] which is more prevalent in some countries than others [46, 31]. As these patterns and correlations might be driven by confounding factors such as organization size or market structure, further investigations are warranted.

For both **closeness centrality** and **degree centrality**, supplier features show a positive correlation with SHAP values, while the corresponding buyer features show a negative or flat relationship (note that these features are calculated on their respective network projections). The positive association of supplier closeness centrality with SHAP values suggests that likely fraudulent contracts are awarded to firms that, on average, have fewer buyer intermediaries with other suppliers (i.e. there are fewer buyers located in between the supplier and the other suppliers in the network). Conversely, for buyer closeness centrality, the association suggests that buyers who are structurally further from other buyers through supplier intermediaries, are more likely to award likely fraudulent contracts. The positive association between supplier degree centrality and SHAP values mean that suppliers that share common buyers with a large number of other suppliers are more likely to receive high risk contracts. While, in the buyer projection, the lack of association with SHAP values may instead indicate that the number of shared suppliers with other buyers is of lesser importance for fraud and corruption risks.

Taken together, the closeness and degree centrality features further strengthen the argument that those suppliers which are central in the network tend to display higher risks of fraud and corruption. This points at potential state capture relationships [12] while also raise the possibility that suppliers dominate in captor networks. Moreover, the negative association of SHAP values with buyer closeness centrality suggests that buyers who are further out in the network, and hence have fewer suppliers who link them directly to other buyers may take advantage of more limited monitoring and hence award higher risk contracts [24].

These relationships also reveal the deeper nature of the labels used by the model: since the labels are derived from company-level sanctions, and contract characteristics tend to be highly similar within the same supplier, the model likely relies more on supplier-related attributes to identify risky contracts. Furthermore, because sanctions are imposed at the company level, the model is better fit to capture contract features that reflect company behavior rather than buyer-side dynamics.

Nevertheless, these are simple explanations for a range of complex relationships without being able to go into great depth. There are certainly a number of alternative explanations, for example, revolving around buyer-supplier size correlation, and market structure. Hence, further analysis could build on these findings and explore them further.

Domain-knowledge features

In Figure 6, we present four continuous domain-knowledge features of the model. Two features display a complex and weak relationship with predicted risks (Supplier Prop. of Recorded Direct Procedures and Supplier Avg. CRI), while the two others behave closer to theoretical expectations (Supplier Number of Contracts and Buyer-Supplier Spending per Active Week).

The complex relationship between SHAP values and both **Supplier Prop. of Recorded Direct Procedures** and **Supplier Avg. CRI** hides two different dynamics, centered around the use of direct awards. Although best practices in public procurement recommend avoiding direct procedures –those contracts that are assigned without competition–, in Mexico these account for 74.8% of all contracts in our dataset. This large subsample in our dataset, however, introduces a likely bias in the labels we use for training, as audits leading to sanctions are more common in non-direct, open procedures, than in direct procedures. This is due to the fact that complex open procedures with extensive audit trail are more readily amenable to audits leading to sanctions rather than simple, direct awards [18]. Hence, it is possible that suppliers which receive exclusively or nearly exclusively direct awards are less likely to be sanctioned due to the lack of auditable information. While among suppliers with low-to-medium direct award share, the expected positive relationships between direct awards as a red flag and also the other red flags are present.

Supporting this interpretation, the SHAP dependence plot of Supplier Avg. CRI (Figure 6) reveals two distinct regions. For feature values below 0.5, 78% of non-direct procedures fall in this range and are associated with higher average SHAP values, although the Pearson correlation is close to zero. For values above 0.5, 93% of real-direct procedures fall in this region, which exhibits consistently low SHAP

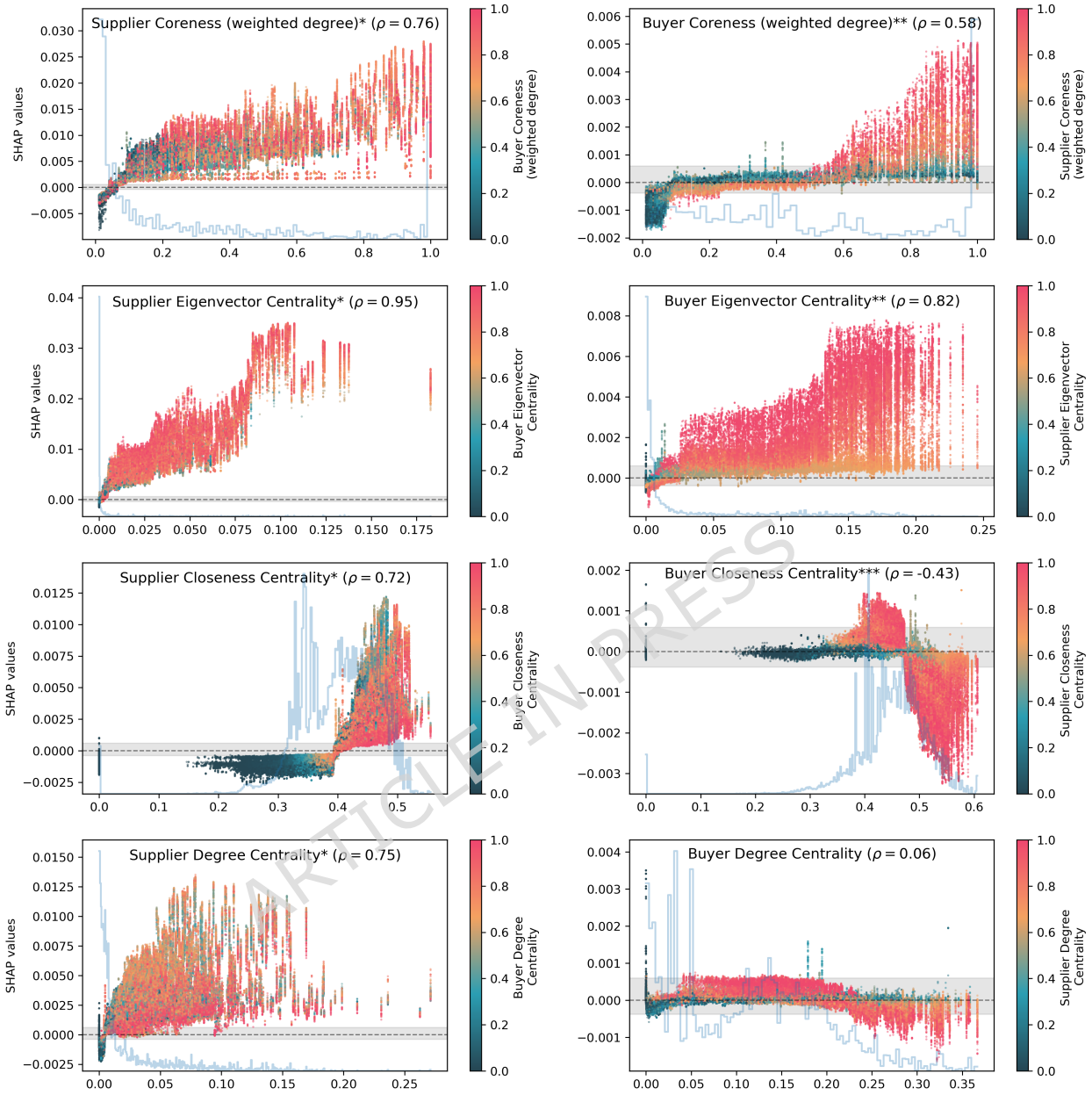

Figure 5: Selected network features' SHAP dependence plots. This figure presents the SHAP dependence plots for ten selected network features from the model and its Pearson correlation(ρ), along with the distribution of the feature in the background (in light blue). The y-axis indicates the features' contribution to the predicted probability of being fraudulent (SHAP value) for each contract, while the x-axis indicates the original values of the feature. The plots in the left column correspond to features related to the supplier, while those on the right correspond to the buyer. Colors correspond to the values of other variables included in the model label. The $*, **, ***$ symbolize the belonging of the feature to the top 10, 20 and 30 most important features of the model,respectively. The grey band around zero represents the 10th and 90th quantiles of the SHAP value distribution across all features, and the blue histogram in the background the distribution of the feature analyzed in the given figure.

values and likewise a weak Pearson correlation. The overall negative correlation is therefore driven by the contrast between non-direct and real-direct procedures, while correlations within each region are small. See also Supplementary Information I.2.

A similar pattern can be found in Supplier Prop. of Recorded Direct Procedures, where we observe a positive correlation with its SHAP values, but only below a specific threshold. For example, for suppliers with less than half of direct procedures, the Pearson correlation is 0.55, while for those with less than 0.9 the correlation is 0.36; and for those with more the 0.9 the correlation turns into negative. This indicates that a higher proportion of direct procedures contributes to a higher predicted probability of fraud, but only for those suppliers whose majority of contracts are non-direct.

Supplier Number of Contracts shows a clear positive relationship with the SHAP values, but only when the proportion of Recorded Direct Procedures is not at its highest level. It is important to note that, although the number of contracts is included as a feature in the model, suppliers within the top 5% in terms of contract count were undersampled in the training set. Consequently, the SHAP value distribution is not biased by the presence of these highly active suppliers. Furthermore, the plot indicates that suppliers with a large number of contracts generally have higher positive contributions to the predicted probability of fraud; however, this relationship holds primarily for suppliers with a relatively low proportion of Recorded Direct Procedures. This further supports the idea that the model could potentially underestimate risks for suppliers with overwhelming non-competitive tenders.

Buyer–Supplier Spending per Active Week shows a positive correlation with SHAP values, indicating that higher spending concentration over shorter time spans increases the predicted probability of fraud. This variable divides the total spending between a buyer–supplier pair by the number of active weeks in which at least one transaction occurred. When analyzed in interaction with Supplier Coreness, it becomes clear that not all suppliers in the network core exhibit high spending per active week. However, those that combine both medium-high and high coreness and high spending concentration show the largest SHAP values, suggesting that temporal concentration of spending is an important risk factor among core suppliers.

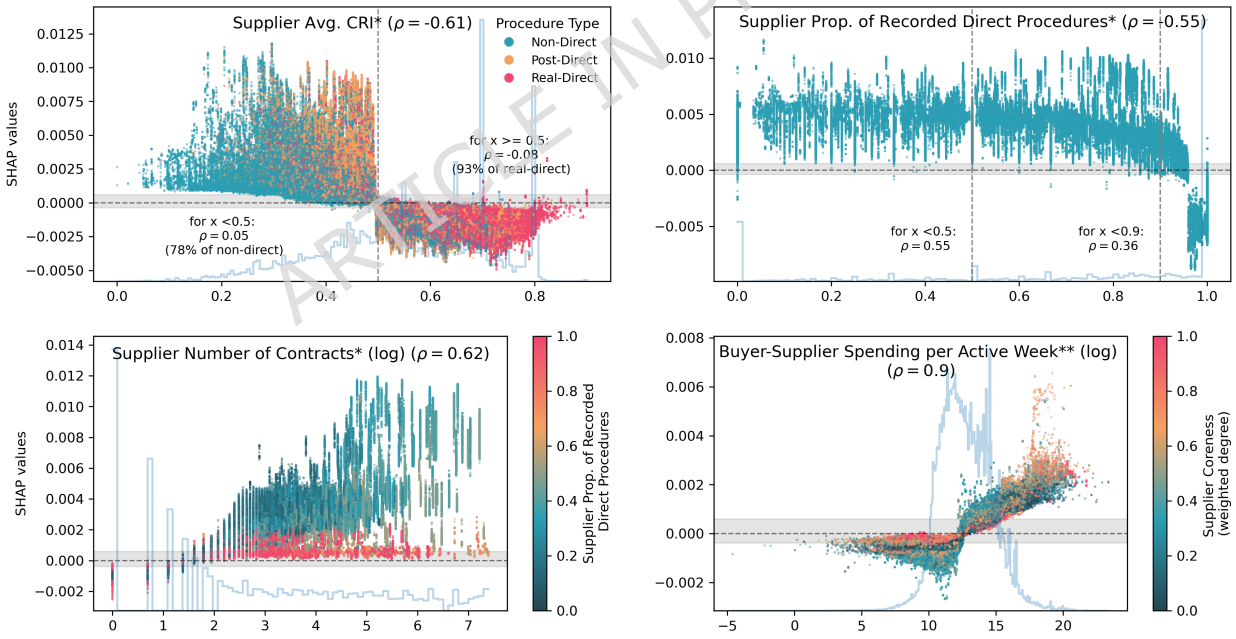

Figure 6: Selected Domain-Knowledge features' SHAP dependence plots. This figure shows the SHAP dependence plots of four selected domain-knowledge features of the model, along with the distribution of each feature in the background and the Pearson correlation (ρ). Colors correspond to the values of other variables included in the model label. The $*, **, ***$ symbolize the belonging of the feature to the top 10, 20 and 30 most important features of the model, respectively. The grey band around zero represents the 10th and 90th quantiles of the SHAP value distribution across all features, the blue histogram in the background the distribution of the feature analyzed in the given figure, and the dashed vertical lines to specific thresholds of features' values.

Conclusion and limitations

In this paper we presented a methodology based on Positive-Unlabeled (PU) learning to detect fraud and corruption in Mexican public procurement using labels based on readily available sanctions information of government suppliers. The developed model is comprehensive, incorporating traditional red flags such as the use of direct awards and also relational information embedded in contracting networks, such as supplier centrality.

We demonstrate that suppliers' sanctions can serve as labels to infer risk factors from, outlining a methodology applicable across a wide array of countries. Nevertheless, our approach faces difficult methodological challenges: we only have positive labels, no negative ones; and the observed labels are very rare in the overall dataset. We used two different algorithms to tackle these problems: PU Bagging [30] and Hellinger Distance Stratified Random Forest (HDSRF) [36] and evaluated their performance based on their average gain and average lift. We observed that the HDSRF model has shown consistently better performance than the other algorithm in the evaluation measures we used and in both transductive and inductive learning. Moreover we also compared ML models to ranking done by the Corruption Risk Index (CRI[13]), which was also surpassed by the HDSRF.

In addition to the performance evaluation, we also unpacked the best-performing model, pointing at relevant variable importances and relationships between fraud risk and risk features. On the whole, network features contribute more to model prediction than traditional red flag indicators, suggesting that the literature should focus more on network methods and features [29]. Furthermore, established red flags contribute to prediction accuracy most when aggregated to the supplier level, rather than on the contract-level.

We provided an interpretation of our best model by presenting the most important features and examining their relationships with the predicted risk score of the model. We used the SHapley Additive exPlanaitions (SHAP values) [27, 26] for interpretation. Regarding network features, from the suppliers' perspective, higher contributions to the predicted probability of a fraudulent contract are associated with (1) being in the core of the bipartite network, (2) being connected to influential buyers and suppliers, (3) have fewer buyer intermediaries to other suppliers, and (4) have a large number of suppliers sharing common buyers. From the buyers' side, the high positive contributions to the predicted probability tend to involve buyers that (1) are present in the core of the bipartite network, and (2) are linked to other well-connected buyers; while negative contributions are associated with (3) buyers with fewer intermediaries. These point at the potential central organizing force of corrupt relationships in the contracting network, observed in a number of European countries [46].

Regarding domain-knowledge features, that is established red flags of corruption, we found a complex, i.e. non-linear, and often weak association with predicted risks of fraud. This aligns with recent literature testing red flags against proven cases, e.g. in Italy by [7]. These findings point at two crucial arguments from prior literature. First, available labels derived from proven cases of corruption, fraud, and related criminal charges are by no means unbiased samples of corrupt behaviors. In some countries, political influence over law enforcement and courts means that some well-connected corrupt actor never get charged while others are charged predominantly for political gain, rather than reflecting the magnitude of their wrongdoing. Yet, in other countries, administrative and legal constraints mean that some types of corruption go unprosecuted. For the latter, we found evidence in Mexico as nearly 75% of contracts in our dataset follow direct procedure types which are less amenable to audit and investigations due to weaker audit trail [18]. For suppliers with substantial non-direct awards (i.e. more than half of contracts), red flags, such as direct award share, have the expected positive relationship with predicted risks.

Second, red flags have been developed to proxy high-level institutionalized forms of corruption (e.g. [9] rather than outright illegal and sanctionable activities. This means that many, yet not all, activities flagged by established red flags are formally legal as high-level political influence assures formal procedural correctness. This argument is supported by literature pointing at the strong association between red flags and personal political connections of suppliers (e.g. [39]), supplier donations (e.g. [44]), and political control over the bureaucracy (e.g. [10]). Based on these two main arguments, it appears beneficial to combine ML-based and traditional red flags and use them in tandem for a more comprehensive measurement. While more research is needed for exploring their alignment as well as their strengths and weaknesses.

Our methodology has some limitations. The first one concerns our labeling strategy. We showed that contracts of the same supplier are very similar to each other and there is little change in supplier contracting behavior before and after a sanction. Labeling all contracts of a sanctioned supplier as positive produces a comprehensive dataset capturing as wide as possible set of corrupt behaviors. Yet, it

is a strong assumption which can induce bias in our model by introducing false positives in the labels for training. Another limitation is that the chosen model, the HDSRF, relies on the SCAR assumption of positive labels. This means that the model assumes that the observed positives are a random selection of the entire universe of fraudulent contracts, which is most likely not true. This can also introduce bias in our results since it can only detect contracts that follow the same corruption strategies as the observed cases, but not innovative or harder-to-investigate fraudulent practices that have not been identified by the authorities.

In terms of generalization to other contexts and deployment in law enforcement practice, our methodology requires several adaptations. First, the model should be trained on contracts within the relevant national context, as the Mexican federal procurement system exhibits specific characteristics that may not apply to other countries and territories. For example, in order to be applied in other countries, it is needed their procurement system to be recorded at the contract-level, with characteristics about the contract value, award date, buyer information, supplier information, but also at the tender level, for example when was the opening and closing date for submission, date of contract award, and at least a significant number of tenders with information about the number of bidders, since these are used to create the red flags and the aggregated CRI indicators. Moreover, it is also necessary for the authorities to keep a standardized record of sanctioned suppliers that can be matched with the contracts dataset.

Second, if our model is deployed for regular contract monitoring and investigations support, it is necessarily that the unseen dataset is of an entire year, since many of our features, including risk indicators and network features, are annually based. Predicting a non-annually dataset would introduce biases in the predictions.

Third, our labeling strategy, which classifies all contracts of a sanctioned company as fraudulent, leads to a valid predictive model under assumptions that sanctions do not alter company behavior over time and that contracts from sanctioned companies are highly similar across different years. If these conditions do not hold, alternative labeling strategies should be adopted.

Fourth, the strategy of under-sampling contracts from large companies in our training and test was adopted due to the high concentration of contracts by supplier which impacted the independence assumption of our machine learning models. However, this strategy is only appropriate when there's strong similarity between contracts of same supplier, specially for large companies. If there is no such similarity, another strategy should be adapted.

However, if the similarity between contracts is fulfilled, the company-based under-sampling train-test split can still be used for deployment in realistic scenarios and in other contexts, since the under-sampling only affects the training set but not the set on which the predictions are made. For example, a logical strategy to use for authorities in the transductive setting would be to split the entire contracts in two subsets, so that one is used as training and other as prediction set, and the other way around. Getting the average of different splits of subsets would be advisable. In the case of the inductive setting, the training set would be the complete available contracts dataset.

Moreover, once the predictions are set and there's a reliable set of contracts with high predicted probabilities, the SHAP values can give guidance for authorities to target specific characteristics of contracts.

In terms of policy implications, our machine learning approach is replicable under certain conditions and it is transparently able to detect likely fraudulent contracts based on domain-knowledge and network features, leveraging publicly available information. In practice, it enables the ranking of contracts and the identification of those with the highest predicted probability, providing a concrete and interpretable tool to guide auditing and investigative efforts. We believe this methodology would be useful for Mexican and potentially other auditing and law enforcement authorities by allowing them to prioritize their efforts to investigate fraud and corruption in public procurement. Moreover, aggregating the predicted risks to higher level units such as regions, markets or years, allows for tracking risk trends over time and can lead to data-informed policy recommendations.

References

[1] Aldana, A., Falcón-Cortés, A., and Larralde, H. A machine learning model to identify corruption in México's public procurement contracts, Dec. 2022.

[2] Almendra, V. Finding the needle: A risk-based ranking of product listings at online auction sites for non-delivery fraud prediction. *Expert Systems with Applications 40*, 12 (Sept. 2013), 4805–4811.

[3] Berrar, D. Performance Measures for Binary Classification. In *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier, 2019, pp. 546–560.

[4] CompraNet. Datos Abiertos de las Unidades Compradoras. http://www.gob.mx/compranet/documentos/datos-abiertos-250375.

[5] Csardi, G., and Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Systems* (2006), 1695.

[6] Czibik, Á., Fazekas, M., Hernandez Sanchez, A., and Wachs, J. Networked Corruption Risks in European Defense Procurement. In *Corruption Networks: Concepts and Applications*, O. M. Granados and J. R. Nicolás-Carlock, Eds., Understanding Complex Systems. Springer International Publishing, Cham, 2021.

[7] Decarolis, F., and Giorgiantonio, C. Corruption red flags in public procurement: New evidence from Italian calls for tenders. *EPJ Data Science 11*, 1 (Dec. 2022), 1–38.

[8] Falcón-Cortés, A., Aldana, A., and Larralde, H. Practices of public procurement and the risk of corrupt behavior before and after the government transition in México. *EPJ Data Science 11*, 1 (Dec. 2022), 1–26.

[9] Fazekas, M., Cingolani, L., and TTth, B. A Comprehensive Review of Objective Corruption Proxies in Public Procurement: Risky Actors, Transactions, and Vehicles of Rent Extraction. *SSRN Electronic Journal* (2016).

[10] Fazekas, M., Ferrali, R., and Wachs, J. Agency Independence, Campaign Contributions, and Favoritism in US Federal Government Contracting. *Journal of Public Administration Research and Theory 33*, 2 (Apr. 2023), 262–278.

[11] Fazekas, M., Tóth, B., Abdou, A., and Al-Shaibani, A. Global Contract-level Public Procurement Dataset. *Data in Brief 54* (June 2024), 110412.

[12] Fazekas, M., and Tóth, I. J. From Corruption to State Capture: A New Analytical Framework with Empirical Applications from Hungary. *Political Research Quarterly 69*, 2 (June 2016), 320–334.

[13] Fazekas, M., Tóth, I. J., and King, L. P. An Objective Corruption Risk Index Using Public Procurement Data. *European Journal on Criminal Policy and Research 22*, 3 (Sept. 2016), 369–397.

[14] Fazekas, M., and Wachs, J. Corruption and the Network Structure of Public Contracting Markets across Government Change. *Politics and Governance 8*, 2 (May 2020), 153–166.

[15] Ferwerda, J., Deleanu, I., and Unger, B. Corruption in Public Procurement: Finding the Right Indicators. *European Journal on Criminal Policy and Research 23*, 2 (June 2017), 245–267.

[16] Galli, S. *Python Feature Engineering Cookbook: Over 70 Recipes for Creating, Engineering, and Transforming Features to Build Machine Learning Models*. Packt, Birmingham Mumbai, 2020.

[17] Ganguly, S., and Sadaoui, S. Online Detection of Shill Bidding Fraud Based on Machine Learning Techniques. In *Recent Trends and Future Technology in Applied Intelligence* (Cham, 2018), M. Mouhoub, S. Sadaoui, O. Ait Mohamed, and M. Ali, Eds., Lecture Notes in Computer Science, Springer International Publishing, pp. 303–314.

[18] Gerardino, M. P., Litschig, S., and Pomeranz, D. Distortion by Audit: Evidence from Public Procurement. *American Economic Journal: Applied Economics 16*, 4 (Oct. 2024), 71–108.

[19] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems 22*, 1 (Jan. 2004), 5–53.

[20] Instituto Mexicano para la Competitividad, and Mexico Evalua. Anexo Metodológico: Mapeando la Corrupción. Tech. rep., IMCO, 2019.

[21] Jain, S., White, M., and Radivojac, P. Recovering True Classifier Performance in Positive-Unlabeled Learning. *Proceedings of the AAAI Conference on Artificial Intelligence 31*, 1 (Feb. 2017).

[22] Jaskie, K., and Spanias, A. Positive And Unlabeled Learning Algorithms And Applications: A Survey. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (PATRAS, Greece, July 2019), IEEE, pp. 1–8.

[23] Jaskie, K., and Spanias, A. *Positive Unlabeled Learning*. No. 51 in Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2022.

[24] Luna-Pla, I., and Nicolás-Carlock, J. R. Corruption and complexity: A scientific framework for the analysis of corruption networks. *Applied Network Science 5*, 1 (Feb. 2020), 13.

[25] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence 2*, 1 (Jan. 2020), 56–67.

[26] Lundberg, S. M., Erion, G. G., and Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles, Mar. 2019.

[27] Lundberg, S. M., and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, Dec. 2017), NIPS'17, Curran Associates Inc., pp. 4768–4777.

[28] Luo, J., Poursafaei, F., and Liu, X. Towards Improved Illicit Node Detection with Positive-Unlabelled Learning, Mar. 2023.

[29] Lyra, M. S., Damásio, B., Pinheiro, F. L., and Bacao, F. Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods. *Applied Network Science 7*, 1 (Dec. 2022), 83.

[30] Mordelet, F., and Vert, J. P. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters 37* (Feb. 2014), 201–209.

[31] Nicolás-Carlock, J. R., and Luna-Pla, I. Organized crime behavior of shell-company networks in procurement: Prevention insights for policy and reform. *Trends in Organized Crime* (July 2023).

[32] Nikpour, B., Rahmati, F., Mirzaei, B., and Nezamabadi-pour, H. A comprehensive review on data-level methods for imbalanced data classification. *Expert Systems with Applications 295* (Jan. 2026), 128920.

[33] OECD. Public Procurement. https://www.oecd.org/governance/ethics/public-procurement.htm.

[34] OECD. *Integrity in Public Procurement: Good Practice from A to Z*. OECD, Apr. 2007.

[35] OECD. Preventing Corruption in Public Procurement. Tech. rep., Organization for Economic Cooperation and Development, 2016.

[36] Ortega Vázquez, C., vanden Broucke, S., and De Weerdt, J. Hellinger distance decision trees for PU learning in imbalanced data sets. *Machine Learning* (Mar. 2023).

[37] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*, 85 (2011), 2825–2830.

[38] ProAct. The Procurement Anticorruption and Transparency platform (ProACT). https://www.procurementintegrity.org/about.

[39] Romero, D. Bureaucratic Capacity and Political Favoritism in Public Procurement. *Comparative Political Studies 58*, 6 (May 2025), 1067–1100.

[40] Saunders, J. D., Alex, and Freitas, A. Evaluating the Predictive Performance of Positive-Unlabelled Classifiers: A brief critical review and practical recommendations for improvement, June 2022.

[41] Secretaría de la Función Pública. Proveedores y contratistas sancionados. https://datos.gob.mx/busca/dataset/proveedores-y-contratistas-sancionados.

[42] Servicio de Administración Tributaria. Listado de contribuyentes (Artículo 69-B del Código Fiscal de la Federación). http://omawww.sat.gob.mx/cifras_sat/Paginas/datos/vinculo.html?page=ListCompleta69B.html, Feb. 2023.

[43] Shapley, L. S. A Value for n-Person Games. In *Contributions to the Theory of Games, Volume II*, H. W. Kuhn and A. W. Tucker, Eds. Princeton University Press, Mar. 2016, pp. 307–318.

[44] Titl, V., De Witte, K., and Geys, B. Political donations, public procurement and government efficiency. *World Development 148* (Dec. 2021), 105666.

[45] Titl, V., Mazrekaj, D., and Schiltz, F. Identifying Politically Connected Firms: A Machine Learning Approach. *Oxford Bulletin of Economics and Statistics 86*, 1 (2024), 137–155.

[46] Wachs, J., Fazekas, M., and Kertész, J. Corruption risk in contracting markets: A network science perspective. *International Journal of Data Science and Analytics 12*, 1 (June 2021), 45–60.

[47] Wachs, J., and Kertész, J. A network approach to cartel detection in public auction markets. *Scientific Reports 9*, 1 (July 2019), 10818.

[48] Wiehen, M., and Transparency International, Eds. *Handbook for Curbing Corruption in Public Procurement*. Transparency International, Berlin, Germany, 2006.

Author Contributions

M.M.H., J.K. and M.F. conceived the idea of the paper and wrote the manuscript. MMH collected the data and implemented the methods.

Data Availability Statement

Data to reproduce our analysis available upon request.

Research Funding

This research was supported by a PhD scholarship granted to Martı Medina-Hernandez by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI, formerly CONACYT) of Mexico. The present article is part of the outcomes derived from this doctoral funding.

Additional Information

Competing Interests: The authors declare no competing interests