Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Preventing rather than punishing: An early warning model of malfeasance in public procurement

Jorge Gallego a,*, Gonzalo Rivero b, Juan Martínez c

a Universidad del Rosario, Colombia
b Westat, United States of America
c Duke University, United States of America

ARTICLE INFO

Keywords:
Public procurement
Corruption
Inefficiency
Machine learning
Forecasting

ABSTRACT

Is it possible to predict malfeasance in public procurement? With the proliferation of e-procurement systems in the public sector, anti-corruption agencies and watchdog organizations have access to valuable sources of information with which to identify transactions that are likely to become troublesome and why. In this article, we discuss the promises and challenges of using machine learning models to predict inefficiency and corruption in public procurement. We illustrate this approach with a dataset with more than two million public procurement contracts in Colombia. We trained machine learning models to predict which of them will result in corruption investigations, a breach of contract, or implementation inefficiencies. We then discuss how our models can help practitioners better understand the drivers of corruption and inefficiency in public procurement. Our approach will be useful to governments interested in exploiting large administrative datasets to improve the provision of public goods, and it highlights some of the tradeoffs and challenges that they might face throughout this process.

© 2020 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

1. Introduction

The “Division of Structured Operations” was the name of the department created by the Brazilian construction company, Odebrecht, to deal with politicians in the countries in which the company operated. A scandal involving Odebrecht and several governments throughout Latin America broke in the mid-2010s and revealed that the division acted as the bribery arm of the corporation. As the former CEO Marcelo Odebrecht confessed, the firm had paid millions of dollars over the years to politicians in countries like Brazil, Colombia, Peru, Venezuela, and Panama in exchange for favorable conditions in public procurement. This example, and many other similar scandals around the world, underscore the significance of malfeasance in public procurement, which has been identified as a roadblock to economic and political development (Bardhan, 2006; DiRienzo et al., 2007; Mauro, 1995; Rose-Ackerman, 1999).

Researchers have suggested several measures to curb corruption, among which transparency of rules, laws, and transactions are common recommendations (Ades & Tella, 1999; Treisman, 2000; Wei, 2000; World Bank, 2013). In the last few years, several countries have adopted web-based platforms for public procurement, wherein agencies are required to register all the requests for the provision of goods or services. However, the volume and complexity of the information stored in these systems often requires specialized skills to identify patterns that may point to wrongdoing. Machine learning tools seem particularly suitable for this purpose. In fact, existing research has already explored this possibility and has tried to predict malfeasance in international development contracts (Grace et al., 2016),...
or within provinces (Lopez-Iturriaga & Sanz, 2018) or counties (Colonelli et al., 2019) in a single country.

We took this research program one step further and developed models to identify procurement contracts that are likely to result in either malfeasance or some other undesirable outcome. To this end, we used data from Colombia’s e-procurement system—the *Sistema Electrónico de Contratación Pública* (SECOPI)—which contains more than two million contracts initiated between 2011 and 2015. We combined this data with lists of vendors that have been reported to or investigated by anti-corruption agencies or have been fined for breach of contract. Our models show the probability that an individual contract will result, down the line, in troublesome outcomes related to corruption, potential violations of the conditions of the contract, or operational inefficiency. We believe that our approach will be useful to government agencies in two different ways. First, the predicted values of the models can be taken as risk scores that anti-corruption agencies, watchdog organizations, and different agencies (in their role as contracting parties) can use to guide the selection of procurement operations that need to be monitored more closely or even audited. In this regard, the tool can be seen as a method to improve oversight of public procurement, reduce redundancy, and detect, as early as possible, situations that will result in economic losses for the government. Second, understanding which variables are more likely to be associated with corruption, breach of contract, and inefficiency can be useful when evaluating potential reforms to regulations and institutions, which is a challenging task when there are potentially hundreds or even thousands of predictors.

Our research is part of a growing body of literature on exploiting new sources of administrative data to improve the ability of governments to fight against corruption and reduce inefficiency (Anderson, 2009; Berton et al., 2010; Prasad & Shivarajan, 2015; West, 2004) and on the application of machine learning tools to better understand areas of the public sphere, such as security (Mena, 2011), education (Kotsiantis, 2012), health (Kleinberg et al., 2015), conflict (Muchlinski et al., 2016), peace (Gallego et al., 2019), and justice (Kleinberg et al., 2018). Although quantitative policy research has benefited enormously from the adoption of causal inferences for program evaluation—that is, to understand whether policy X has a causal impact on outcome Y—we maintain that in some situations, predictions can be more useful than causal understandings (Kleinberg et al., 2015). Being able to anticipate whether a given outcome will occur is crucial for governments who need to allocate scarce resources across competing priorities often surrounded by uncertainty. The early detection of fraud, malfeasance, or inefficiency is a perfect example of an area in which the use of a predictive approach can help optimize the provision of public goods.

Our models exhibit high predictive performance, although, as we discuss below, this evaluation is largely dependent on the practical applications and resource constraints of the agencies using them. In some cases, governments may prefer more aggressive classifiers that minimize the expected number of false negatives—for instance, to scrutinize suspicious contracts that are not predicted as such by the model—to cast as wide a net as possible. However, given the availability of resources for monitoring, governments can reasonably prefer to minimize false positives and focus their attention on cases that are more likely to result in irregularities. We discuss different metrics to evaluate the models in the context of their use.

Despite the large number of features to which we had access (more than 300), our study showed that only a small number of characteristics have a significant influence on the detection of malfeasance, breach of contract, and inefficiency. We found that the typical characteristics of contracts, such as their size (measured by their estimated budget and duration), delays in their implementation, time before the next election, and geographical and sector-specific patterns, are some of the key features that matter the most across our specifications. As a result, a combination of the main traits of these projects, along with some political attributes associated with electoral cycles, are potential predictors that can be used by public agencies to target contracts that warrant closer monitoring.

The remainder of the paper is structured as follows. In Section 2, we discuss some of the theoretical foundations that underpin our approach to the use of machine learning in the context of early warning methods for detecting corruption, including some background information about the adoption of e-procurement platforms in the case of Colombia. Section 3 describes the data used in this study, which comprise public procurement information, evidence of malfeasance or breach of contract, and predictors at the municipal level. In Section 4, we describe the models we developed for the analysis, as well as some particularities, such as the fact that the number of cases with malfeasance or inefficiency represented only a small minority of the total contracts registered in SECOPI. The results of the analysis, as well as an interpretation of the models, are presented in Section 5. We offer conclusions and directions for future research in Section 6.

2. E-procurement and the control of inefficiency and corruption

2.1. Inefficiency and corruption

Inefficiency and corruption are pervasive problems in the developing world. Troves of literature discuss the

---

1 We measured inefficiency through extensions to the original contract in terms of either time or money. The rationale for this variable is discussed in Section 3.

2 It is important to keep in mind that practitioners need to interpret these results with care, as the approach we undertake is not causal in nature. To fully address the problem of causality, alternative approaches such as randomized controlled trials and quasi-experimental techniques are necessary. We see our models as a complement to these approaches. For excellent discussions on the connections between causal inference and machine learning, we refer the reader to Belloni et al. (2017, 2014), and Athey and Imbens (2019).

3 We define corruption as the use of public office for private gain (Rose-Ackerman 1999).
negative impact of these phenomena on many important outcomes such as growth (Mauro, 1995), the quality of democracy (Mungiu-Pippidi, 2015), and service delivery in terms of security (Condra et al., 2019), education (Dufflo et al., 2012), and health (Chaudhury et al., 2006). As noted in the seminal study of Bandiera et al. (2009), corruption, especially in the context of public procurement—the process through which the government acquires goods and services—is only the active version of a more general problem in which the government is the damaged party in a transaction. In other words, while corruption refers to a situation in which public servants interfere in the process of public service delivery for personal benefit, the problem that governments face in procurement has to do more generally with distortions induced by any actions from public servants or vendors, regardless of whether they aim at personal gain.

Thus, we can see corruption as one extreme on a spectrum of situations where the government suffers economic losses. This includes not only extortion, bribery, solicitation, favoritism, nepotism, and an array of other forms of corruption, but also agreements made with a vendor where the latter does not perform according to the original parameters of the contract or does not perform at all. In other words, it includes situations in which the vendor either does not deliver at the cost or within the timeline originally stipulated or does not produce the good or service that was contracted. We can then think of corruption, breaches of contract, or operational inefficiencies by vendors as different flavors of violations of the expectations of a transaction (Colombatto et al., 2001). There are, of course, important differences between the legal implications of active waste implied by corruption and the passive waste (Bandiera et al., 2009; Lagunes, 2017) induced by vendors failing to provide what was expected of them. In any case, it is reasonable to expect that governments will be interested in preventing both types of situations and ensuring that the procurement process does not open the door to intentional or accidental misbehavior.

Furthermore, government agencies have a direct motivation to prevent these troublesome outcomes. The literature shows that these phenomena (active and passive waste) are often empirically interrelated. For example, Dal Bo and Rossi (2007) and Bosio et al. (2020) found a positive correlation between corruption and inefficiency. Because these distortions could be the result of rent-seeking and opportunistic behavior, it is then reasonable to expect that outcomes such as breaches of contract or deviations from the conditions of a contract can be taken as potential warnings of corruption.

While corruption is hard to detect, numerous studies show that it is common in public procurement (Bosio et al., 2020; Campos et al., 2019; Collier & Kirchberger, 2016; Colonnelli et al., 2019; Colonnelli & Prem, 2020; Decarolis, 2014; Decarolis & Palumbo, 2015; Di Tella & Schargrodsky, 2003; Gallego, Prem et al., 2020; Lichard & Fernandes, 2019; Olken, 2007). Moreover, malfeasance manifests in different forms, including cost overruns (Flyvbjerg et al., 2003; Gallego, Prem et al., 2020; Olken, 2007), favoritism (Baranek & Titl, 2020; Burgess et al., 2015; Mironov & Zhuravskaya, 2018), collusion (Coviello & Gagliarducci, 2017), and bid rigging (Conley & Decarolis, 2016). Given that a sizeable fraction of the provision of public goods is made through public procurement, finding solutions to this issue is crucial for the development of countries.

Importantly, it is possible to link corruption and inefficiency through a common lack of government accountability (Besley, 2006). If we think about the problem as a principal–agent relationship, information asymmetries can be used to explain the prevalence of inefficient outcomes (Ivanov, 2007; Johnston, 2001; Lawson, 2009; Riley, 1998). In other words, regardless of whether we consider it as a relation between governments and voters (Adsera et al., 2003; Besley, 2006; Myerson, 1993), or between elected officials and appointed bureaucrats (Becker & Stigler, 1974; van Rijckeghem & Weder, 2001), lower levels of information and a lack of transparency translate into poor delivery of social services.

A substantial debate in the literature has focused on whether bottom-up accountability mechanisms are more or less effective than top-down strategies to fight corruption. Olken (2007) claimed that corruption and inefficiency are lower when top-down audits are used to supervise projects, as opposed to bottom-up accountability strategies. This underscores the importance of government agencies and watchdog organizations that aim to gather information on the performance of public servants. Other studies, such as Bjorkman and Svensson (2009), Ferraz and Finan (2008), and Chong et al. (2015), showed how information is crucial to curb corruption, as better informed voters are more effectively able to hold elected officials accountable (but see Dunning et al. (2019)). In the end, whether through top-down or bottom-up strategies, information and transparency are key elements associated with good governance (Prasad & Shivaranj, 2015).

In this regard, e-procurement has been shown to have the potential to reduce inefficiency and corruption (Lewis-Faupel et al., 2016) by improving transparency in various contexts (Berton et al., 2010), such as India (Bhatnagar, 2003; World Bank, 2004), Pakistan (Anderson, 2009), Chile (Shim & Eom, 2009), Fiji (Pathak, 2009), and Korea (Lee, 2009), among others. With tools like e-procurement systems and platforms (of which SECOP in Colombia is an example) that have emerged around the world, there are new opportunities for monitoring insofar as anti-corruption agencies and the public have access to granular information about the process through which governments acquire goods and services.

Equipped with these new data, as we show below, we are afforded new methods of fighting corruption and other outcomes of interest for government agencies. Predictive models, such as the ones we employed, are thus valuable because they enable us to create scores that can measure the likelihood that an individual contract will become problematic. In particular, we worked with 3 different outcomes related to investigations of corruption, breaches of contract, and operational inefficiency.
Our approach is useful in four respects. First, by using several outcomes we offer a more comprehensive view that encompasses direct and indirect outcomes of interest to government agencies. Second, although two of our outcomes were measured at the vendor level, the granularity of our data allowed us to produce scores at the contract level—a scale of analysis more useful to many subnational units and governmental agencies, as it can be used to monitor more precisely contracts that are particularly prone to fraud. Determining which contracts are more likely to be involved in fraudulent activities is useful for anti-corruption agencies and watchdog organizations, because audits are expensive and time-consuming, and not all contracts can be monitored and investigated. Third, our models exhibit high performance and can thus be used as an early warning system by officials responsible for those contracts. As such, they can be used to deter potential misbehavior (as opposed to sanctioning it). Thus, in our view, these models can be used at different stages of the procurement process to mitigate the incidence and impact of corruption. Finally, given that these models can indicate which contract- or municipality-level variables are more predictive of malfeasance or inefficiency, they can also be useful to start a discussion around institutional or procedural reforms to curb corruption.

However, there is an important challenge that needs to be discussed: contractors and public officials are not static agents. Rather, they adapt to new conditions. One would expect that if these algorithms start being used by agencies, actors involved in corruption will adjust their behavior in order to reduce the probability of being detected by any model. Consequently, the application of these machine learning tools needs to be adaptive and dynamic as well. Training and test datasets need to be constantly fed with new information, in order to account for changes in behavior on the other side of the relationship.

2.2. E-procurement in Colombia

The Sistema Electrónico para la Contratación Pública (SECOP) is the first attempt by the Colombian government to digitize and enhance the monitoring of public procurement in the country. SECOP’s design rests on the idea that all contracts between government agencies and private vendors should be made public, including request notices and the awards of each contract. The mission of SECOP is thus to shed light on the partnerships in which public entities engage. As a mandatory step in the procurement process, government agencies must record, store, and publish all procurement actions, documents, and changes using the SECOP platform. As a result, SECOP includes offers, contracts, evaluation reports, and requests for information, studies, and any other documents related to the provisions of goods and services to any government agency.

The regulation of public procurement in Colombia gives government agencies the possibility of using different awarding instruments depending on the characteristics of the contract. Available options include public tenders, the abbreviated choice method, selection based on qualifications and merits, direct selection, and the minimum-value contract method. In theory, the most common practice is that the selection of vendors goes through public tender processes unless the context allows something different. In practice, however, government agencies find ways to use other methods that are more prone to malfeasance, such as the direct selection of providers (see Section 3.2).

3. Data

3.1. Predictors

In our analysis, we used data on public procurement in Colombia between 2011 and 2015 from the SECOP database. As of 2015, the data comprised 886,242 contracts, representing an aggregate value of COP $121,255 billion (on the order of USD $37 billion). The dataset is publicly available on the Internet. As a result, our starting dataset contained 2,241,271 observations and 58 variables, with each observation corresponding to a single transaction between a government entity and a vendor of a service or a good.

The SECOP dataset provides information on the different parties involved in a transaction and about the transaction itself. For instance, it indicates whether the agency is a national or a regional entity, and provides administrative information about the vendor. All purchases are identified by the contract’s approval date, the date the execution of the contract began, the duration of the contract specified in the requisition process, and the date the contract was considered to be fully executed. For each contract, the type of process that was used in the procurement and its current status is given (for instance, whether it has been adjudicated or paid), along with a unique standard identifier (UNSPSC code) for the good or service under contract, a short text description of the goal of the contract, the origin of the resources used to carry out the purchasing process, the planned budget, and the value that was finally assigned in the contract. We also controlled for the total number of contracts assigned to the vendor of each of them. In sum, there were two types of variables in the dataset: a majority of features that were observed before the contract was executed, and a

---

5 Public tenders correspond to contracts that are published by a public agency to seek offers from suppliers and private contractors who can provide goods, services, or products that an organization requires, with the decision ultimately being made based on price and quality of the deliverable.
6 This method is intended to be used to contract standardized products or services.
7 Merit selection is a tool to hire consultants and advisors for public entities.
8 Direct selection is only allowed in cases of manifest urgency, lending contracts, or inter-administrative contracts.
9 This method is used in cases where the procurement is below a specified amount.
10 Information is available on the website, Datos Abiertos de Colombia (Open Data from Colombia): http://datos.gov.co.
handful of features observed after the contract had been completed. We focused on the latter set of variables.

In our analyses, we also used information from the Panel Municipal compiled by the Centro de Estudios sobre Desarrollo Económico of Universidad de los Andes (Acevedo & Bornacelly, 2014), a comprehensive dataset with information at the local level that includes the demographic, economic, fiscal, health, education, and electoral characteristics of each municipio. This information was used in some of the analysis below for a subset of contracts that were executed at the local level. This expanded dataset thus added 132 additional covariates to our models.

3.2. Measures of malfeasance and inefficiency

We used three different indicators to capture different types of outcomes associated with malfeasance, breach of contract, and inefficiency.

One of our variables came from the list of vendors that received fines by the Contraloría General de la República (Office of the Comptroller General) or any of its associated regional offices in the departamentos (equivalent to states in the U.S.) or major municipalities. Several agencies in Colombia have a mandate to prevent and punish inefficiency and corruption, particularly in the case of public procurement. Most important among them, the Contraloría is in charge of the fiscal control of the country, and, as such, its main task is to ensure the proper allocation of public funds and resources. Consequently, this agency audits government entities, which may result in warnings and additional investigations. Some of these investigations may, in turn, lead to a full criminal prosecution by the Fiscalía General de la Nación (Office of the Attorney General) against either public servants, private parties, or both. Investigations by the Contraloría can be the outcome of its own regular audits, they can be initiated at the request of other agencies (such as the Fiscalía), or they can be triggered by whistleblowers, including citizens. It is important to note that the information from the Contraloría is available at the vendor level. Thus, it is not contract-specific, which means that our outcome variable is best understood as determining whether a contract was signed with a vendor that was later sanctioned for irregularities in the management of public resources.

Our second indicator came from the Confederación Colombiana de Cámaras de Comercio (Colombian Confederation of Chambers of Commerce), also known as the Confecámaras. The Confecámaras does not independently conduct audits or investigations. However, as part of its professional activities, it collects and makes available to its members a database of vendors that have been fined due to a breach of contract. Thus, unlike the Contraloría, the Confecámaras aggregates information provided by other sources (the municipios) to the regional chambers of commerce. Importantly, this outcome is also measured at the contractor-level and only refers to breach of public contracts. It is interesting to note that most vendors that are listed by the Confecámaras do not appear in the list from the Contraloría (see Table 1), which is further indication that the two variables capture different types of behavior.

Our third outcome variable exploited the fact that SECO provides information on time and monetary extensions to contracts. Extensions to contracts are considered a sign of inefficiency and they have been linked to irregularities in the procurement process. While different in nature from the information provided by the Contraloría or Confecámaras, we argue that it is relevant for government agencies. First, extensions are an indication of so-called passive waste (Bandiera et al., 2009; Lagunes, 2017), which can be as harmful as corruption in that it represents a misallocation of public resources. Second, the literature frequently points out that there is a positive correlation between inefficiency and corruption (Bosio et al., 2020; Dal Bo & Rossi, 2007; Mauro, 1995). In the case of Colombia, it has been shown (Henao & Isaza, 2018) that a common mechanism to extract more resources from the public coffers is to bid below cost or with implausibly short execution times in order to be awarded a contract and then request extensions. As a result, anti-corruption agencies around the world often have a dual mandate to fight both corruption and inefficiency in public procurement (Bandiera et al., 2009; Lagunes, 2017). Consequently, from the perspective of the government, models predicting inefficiency may be useful in themselves and also a first step toward fighting corruption. Finally, information about extensions is available at the contract level and therefore at a finer scale than the outcomes from the Contraloría or Confecámaras.

### Table 1

|                  | In Confecámaras | Not in Confecámaras |
|------------------|----------------|---------------------|
| In Contraloría   | 83             | 38,677              |
| Not in Contraloría | 23,505         | 2,179,006           |

Note that the number of “positive” contracts in the Confecámaras (23,505) was lower than those in the Contraloría (38,677). This result underscores that breached contracts are not a subset of corrupt contracts detected by the Contraloría. In fact, given that the sanctions and fines reported by the Confecámaras are imposed by the procurers, it is unsurprising that there is little overlap between the two measures or that the Contraloría is more active in detecting irregularities. It stands to reason that corrupt public servants will not sanction their complicit providers. Such behavior can only be detected by an independent anti-corruption agency.

For each case, we had information on the original planned duration (and budget) of the contract, and its actual duration (and budget). We used the former as a predictor variable in our models. Any difference between planned and actual values indicates that an extension took place.

---

11 A limitation of our data is that we needed to assume that if a vendor was sanctioned in year \( t + 5 \), for instance, the vendor was classified as “corrupt” throughout the whole period of analysis. Unfortunately, we did not have the exact date of the illicit action. As such, the best way to interpret this outcome is as an indicator of whether the contract was awarded to a vendor prone to committing acts of corruption.
Three notes of caution are relevant regarding how data availability affected our design. Our measures of malfeasance from the Contraloría were prone to what Lakkaraju et al. (2017) called the “selective labeling problem”—the fact that observed outcomes are the result of choices by human decision-makers. For instance, in our case, investigations started by the Contraloría were the result of a decision by the auditors as to what and what not to investigate. As a result, what our models captured was not necessarily the existence of malfeasance but rather malfeasance that had been detected by the agency. In spite of this limitation, our approach provides valuable insights from a practical standpoint. While our models may be learning from labeled outcomes that do not cover all instances of corruption or types of corruption that auditors have not detected in the past, there is still value in identifying patterns in the data similar to those that auditors, in a more costly process, would have detected. Of course, our outcome regarding contract extensions was not affected by this concern.

The second limitation is related to the granularity of the data. The models that used information from the Contraloría and Confecámaras gave us estimates of the risk that a contract will result in malfeasance or a breach. However, the information available is at the vendor level rather than the contract level. Naturally, it would be more useful to have information on investigations at the contract level to learn about contracts from a vendor with a history of malfeasance that did (or did not) result in wrongdoing. Alternatively, if additional information were available about the vendors themselves, we could instead try to calculate the risk of corruption at the vendor level. Unfortunately, information about vendors is sparse, especially for contracts executed by individuals. We believe that our compromise between contract-level predictors and vendor-level outcomes still offers valuable insights that can be used to allocate anti-corruption resources efficiently. Finally, as noted above, the information on extensions was reported at the contract level.

Finally, it is worth discussing how timely our predictions are. As we argued above, the variables used to feed the classifiers were all measured before the contracts were completed. As such, the tool may be used by anti-corruption bodies and watchdog organizations aiming to prevent malfeasance while projects are ongoing. In the case of procurers, the tool may also be used to optimize resources to oversee contracts after they have been awarded. One might ask, however, whether it is possible to use the tool before a contract has been awarded. This is a tricky issue, as many of the SECOP contract-level variables are in fact the result of the decisions of procurers, who may, in turn, be interested in committing the acts of corruption themselves. Consequently, we believe that our tool is useful for supervisory bodies. Nonetheless, an interesting avenue for future research would be to extend this type of model to create early warning systems at the contractor level. This would be useful to public authorities who wish to anticipate the risks of hiring certain contractors.

4. Models

In order to predict whether a contract will result in malfeasance, breach of contract, or inefficiency, we used two common machine learning models: a lasso classification model (Tibshirani, 1996), and a gradient boosting classification model (GBM) (Friedman, 2001). These two approaches bookend the range of choices between interpretability and flexibility that is available to policymakers when it comes to using modern statistical learning tools for risk scoring. On the one hand, lasso regression is a natural generalization of the well-known logistic regression that preserves the linearity in coefficients and adds a penalization on the sum of the coefficients to control model complexity (Tibshirani, 1996), which greatly simplifies variable selection. Such a simple strategy has proven successful in a variety of domains, but, more importantly, it is easy to fine tune and quickly fitted, even for large datasets, when compared to other standard machine learning methods. In addition, the results can be interpreted in a way that is similar to logistic regression, which is very popular in social science research.

On the other hand, the GBM (Friedman et al., 2001) is common in practical applications in both industry and research and exemplifies well what is commonly known as a black-box model: a model with excellent performance across domains but one that is notoriously difficult to interpret (Breiman, 2001). The GBM belongs to a class of models that combines a number of “weak” learners—in this case, very shallow trees—that are known to underfit the data into a combined “strong” learner with high predictive performance. This particular model builds a potentially large greedy sequence of trees in which, at each step, a classification tree is fitted to data that has been weighted by the residuals from previous steps. In this way, the GBM builds an ensemble of trees that achieves a low generalization error by iteratively upweighting observations that were harder to classify by the preceding history of trees. The model complexity of the GBM is controlled by a number of hyper-parameters that set the maximum length of the sequence of trees (number of trees), the depth of each individual tree (interaction depth), the amount by which each additional step contributes to the final prediction (shrinkage), and the minimum number of observations in each terminal node. As a result, the GBM is more flexible than the lasso, although it requires less intuitive fine-tuning (Chen, Song et al., 2018; Fisher et al., 2018; Friedman, 2001; Molnar, 2019; Ribeiro et al., 2016).

14 We used the implementation in the glnet package (Friedman et al., 2010).
15 We chose the traditional gbm (Ridgeway, 2017) implementation instead of the more popular xgboost (Chen, He et al., 2018) in R because it is better equipped to analyze categorical variables. In particular, the xgboost function expects categorical variables in a one-hot encoding, which means that at the interpretation stage, the user must reconstruct the original categorical variable, even if some categories are not included in the final model.
16 For a description of the full set of hyper-parameters, see Ridgeway (2017).
While many other methods could have been used to model our data,\textsuperscript{17} we settled on lasso and the GBM as they are frequently used in related literature and because they illustrate a common dilemma for analysts between models with higher predictive performance and models whose predictions can be explained to stakeholders downstream.

To fit the models, we took two samples of 225,000 observations from the SECOP database.\textsuperscript{18} One sample was a simple random sample of all the contracts in the database. The second sample was extracted from those contracts in SECOP that were signed at the level of the municipio. We then processed the two subsamples in the same way. We used 75% of the observations for training and kept 25% as hold-out. The model performance reported in the following pages is the hold-out performance. All models were trained using five-fold repeated cross-validation with a grid search.\textsuperscript{19} The grid we used is shown in Table 2. For each combination of values in the grid, we computed the average AUC across the testing folds. The final model complexity was chosen at random among the combinations that yielded an average AUC of no more than one standard deviation away from the highest observed one (see Friedman et al., 2001).

All of our outcomes of interest were relatively infrequent events, i.e., cases with malfeasance, breach of contract, or inefficiency (the “positive class”). They represent a small minority of the contracts (see Table 3). Indeed, more than 95% of the contracts in SECOP were not listed by the Contraloría or Confecámaras, and only around 11% of the contracts received an extension. Such imbalance sometimes, but not always, causes issues with classification tasks (Kuhn & Johnson, 2013) using metrics like accuracy, because models can attain high performance by predicting all instances as belonging to the majority class. In our case, the imbalance was not as severe as in other examples given in the literature. Nevertheless, we tested two popular ways of managing the underrepresentation of “positive” examples.

\begin{table}[h]
\centering
\caption{Values used for the grid search in hyper-parameter optimization.}
\begin{tabular}{llc}
\hline
Model & Parameter Name & Values \\
\hline
Lasso & \text{log}(\lambda) & \{-5, -4.9, -4.8, \ldots, 5\} \\
GBM & Number of trees & 30 regularly-spaced points between 10 and 30,000 \\
& Maximum tree depth & 1, 2, \ldots, 6 \\
& Shrinkage & 0.05, 0.01 \\
& Minimum node size & 25 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Number of observations in each of the outcomes.}
\begin{tabular}{lccc}
\hline
 & Confecámaras & Contraloría & Extension \\
\hline
Negatives & 2,217,692 & 2,202,513 & 1,989,784 \\
Positives & 23,579 & 38,758 & 251,487 \\
\hline
\end{tabular}
\end{table}

First, both lasso and the GBM are suitable for incorporating differential classification costs via weights. Thus, we used a weight that increased the weighted proportion of the positive class to at least 25%. Specifically, we used a weight of 25 for variables collected by the Contraloría and Confecámaras and a weight of 10 for the variable that captured an extension to the contract. The second approach we tested involved resampling to achieve a more balanced distribution of the outcome variable (Van Hulst et al., 2007). Several alternatives are offered in the literature: it is possible to down-sample observations in the most frequent outcome class (Kubat & Matwin, 1997) or, alternatively, the minority class can be up-sampled (Ling & Li, 1998). Popular, modern alternatives combine down- and up-sampling through the creation of synthetic cases (Chawla et al., 2002; He et al., 2008), but these methods often assume that the available features are continuous. Because the majority of the columns in our dataset were categorical, synthetic methods risked causing damage. Thus, in the analysis below we relied on simple up-sampling. It is worth noting that the methods used to address imbalance had no discernible effects, as observed below.

In what follows, we show the results for each of the three outcome variables of interest in four scenarios using different models and methods for dealing with sample imbalance. In addition, for the sample of contracts at the municipal level, we added two additional scenarios in which we used or omitted covariates from the Municipal Panel. In the following section, we discuss model performance by separating the models at the national level (Section 5.1.1) and at the municipal level (Section 5.1.2). The results are interpreted in Section 5.2.

5. Results

5.1. Model fit

5.1.1. All cases

Fig. 1 shows the ROC curves for each of the models using data in the hold-out set. The figure represents the true-positive rate (the proportion of, for instance, detected corruption cases that were correctly classified) and the false-positive rate (the proportion of incorrectly guessed corruption cases) achieved by the classifier for

---

\textsuperscript{17} We also tested support vector machines, extreme gradient boosting, and conditional inference trees, with similar results. Appendix Section E compares the results of a random forests model with those of the GBM, showing that they are similar. The random forests model is not strictly superior.

\textsuperscript{18} Although it would have been desirable to use the entire sample of more than two million contracts, the complexity of the models used (especially the GBM), and the fact that some were estimated multiple times, increased the computational demand of our tasks. In any case, using random samples of 225,000 observations introduced noise to our estimations, if anything.

\textsuperscript{19} We used the caret package in R (Kuhn, 2018), which offers a common interface for a large number of machine learning models as well as utilities for resampling and evaluation.
a sequence of probability cutoff points. As a reminder, the classifier performs better as it approaches the top-left corner of the plot where all cases are correctly classified, and it is indistinguishable from random guessing as it comes closer to the 45-degree line—also displayed in the figure.

The performance of all the models was high by conventional standards, with AUCs ranging from 0.788, where the lasso model predicted extensions, to 0.912, where the GBM predicted inclusion in the Confecámaras list, in both cases using weights to correct the sample imbalance. The models using up-sampling performed slightly worse across the board, with AUCs ranging from 0.785 to 0.897 for the same two models. More importantly, in all cases the GBM performed better than the lasso, regardless of the outcome and the way in which we dealt with outcome imbalance. And the AUC was consistently higher when predicting the Confecámaras outcome than for any of the other two variables.

Researchers (see Cranmer & Desmarais, 2017; Saito & Rehmsmeier, 2015) commonly recommend evaluating sample-imbalanced models using additional tools, like the precision–recall space, which replaces the false-positive rate (positive cases incorrectly predicted as false) with the precision of the model (the proportion of correctly predicted cases over all cases). The precision–recall curves for the same models displayed above are shown in Fig. 2.

Fig. 2 explains a key point inferred from the results in Fig. 1. In all models, the GBM outperformed lasso under all settings. However, the precision–recall curves suggest a different ranking of the models according to their predictive capacity: models predicting inclusion in the Confecámaras list tended to flag more cases than actually existed, whereas the model predicting an investigation by the Contraloría was much more balanced. In addition, contrary to what we see in Fig. 1, the highest performance was achieved by the model that predicted extensions to contracts, regardless of how we dealt with imbalance.20

Figs. 1 and 2 show the performance of the models for all possible ways of turning predicted probabilities into cases that should and should not be investigated. In practice, however, anti-corruption agencies are likely only to have resources to investigate a small subset of all cases that are predicted in the positive class. Therefore, we must determine the extent to which these models can help public servants better target contracts with their limited resources. Rather than measuring the effectiveness of the model at separating cases of potential corruption from others, agencies will benefit more from knowing how often the model will make incorrect predictions of a given number $K$ of contracts with the highest predicted probability of corruption.21

---

20 It is worth noting that this high performance is similar to what was reported in other studies, even under different setups. Colonnelli et al. (2019), for instance, analyzed Brazilian municipalities and obtained maximum levels of AUC, precision, and recall of 0.98, 0.94, and 0.93, respectively. Lopez-Iturriaga and Sanz (2018), used information regarding Spanish provinces and reported accuracy levels ranging from 0.74 to 0.84. Finally, in the case of international development contracts, (Grace et al., 2016) reported an accuracy of 0.7.

21 We thank an anonymous reviewer for this observation and for the recommendation to show the metrics in Table 4.
Some alternative metrics that address this use-case are shown in Table 4. In particular, we show the mean average precision (MAP) of the model for the \( k = 100 \) and \( k = 1,000 \) contracts with the highest predicted probability of malfeasance/inefficiency to show how many of these indeed corresponded to cases that were investigated. Similarly, we also show the normalized discounted cumulative gain (NDCG) for the top \( k = 100 \) or \( k = 1,000 \) contracts to ensure that the cases ranked higher by the model are also more likely to correspond to contracts that were actually investigated. The NDCG calculates the discounted cumulative gain, \( \text{NDCG}_k = \sum_{i=1}^{k} \frac{r_i}{\log(1 + i)} \), which penalizes (by a factor proportional to the logarithm of the order of case \( i \)) true cases of malfeasance/inefficiency that were suggested to agency \( r_i \) if we select the top \( k \) contracts.

With these two metrics, we evaluated the models from the point of view of an agency that can investigate up to \( k \) contracts and that will likely focus on the \( k \) contracts with the highest predicted probability. Given that these metrics target the probabilities produced by the model, we also show the Brier score in the last column of Table 4, as this score can be used to measure the relative aggregate accuracy of predicted probabilities across different methods (Bansak, 2019).

These metrics confirmed what is shown in Fig. 2—that predicting inclusion in Confecámaras is harder than any of the other two outcomes. Of the top 100 contracts ranked by their predicted probability, 25 were indeed listed by the trade board. That proportion fell to 10% for the top 1000 cases. The results were much more positive with the model that predicted contract extensions and investigations by the Contraloria, with over 60% of positive cases found among the top 100 predictions in all cases, and 20% among the top 1000. These results are consistent with the results of the NDCG. It is also significant that all metrics performed better when we up-sampled the data as opposed to weighting the observations. Again, the GBM outperformed lasso in all scenarios.

5.1.2. Municipios

As indicated in Section 4, models at the municipal level can further exploit the availability of the economic, social, and political characteristics of the municipio. This tells us something about the conditions that make investigations more likely.

Fig. 3 shows the ROC curves and Fig. 4 shows the precision–recall curves. The results of using additional covariates from the Municipal Panel in the prediction are shown in red. In black, we represent the models that rely exclusively on predictors available in the SECOP database. The results still show high AUCs within the range of 0.785 and 0.898 but with the ROC overestimating how well we can predict the cases that were included by the Confecámaras relative to an investigation by the Contraloria. As above, the precision–recall curves sorted the performance of the models in the opposite order as the ROC, and the GBM outperformed the lasso.

The most interesting result has to do with the value of using additional data in the predictions. We found a weak effect from using auxiliary data on the performance of the models, with the possible exception of the lasso. This result is consistent with the idea that the lasso imposes a very simple linear structure, and that

![Fig. 2. Precision–recall curves for all models using the three outcome variables.](image-url)
Table 4
Performance of the models.

| Model     | Outcome       | Unbalance | $X_c$ | MAP$_{100}$ | MAP$_{1000}$ | NDCG$_{100}$ | NDCG$_{1000}$ | Brier  |
|-----------|---------------|-----------|-------|-------------|--------------|--------------|---------------|--------|
| GBM       | Confecámaras  | Raw       | No    | 0.23        | 0.11         | 0.64         | 0.57          | 0.05   |
| GBM       | Confecámaras  | Upsample  | No    | 0.28        | 0.11         | 0.66         | 0.57          | 0.01   |
| GBM       | Contraloría   | Raw       | No    | 0.61        | 0.22         | 0.83         | 0.63          | 0.06   |
| GBM       | Contraloría   | Upsample  | No    | 0.91        | 0.23         | 0.97         | 0.65          | 0.01   |
| GBM       | Extension     | Raw       | No    | 0.79        | 0.49         | 0.91         | 0.76          | 0.20   |
| GBM       | Extension     | Upsample  | No    | 0.94        | 0.57         | 0.98         | 0.80          | 0.08   |
| Lasso     | Confecámaras  | Raw       | No    | 0.20        | 0.11         | 0.62         | 0.56          | 0.05   |
| Lasso     | Confecámaras  | Upsample  | No    | 0.22        | 0.12         | 0.64         | 0.57          | 0.01   |
| Lasso     | Contraloría   | Raw       | No    | 0.22        | 0.13         | 0.64         | 0.57          | 0.09   |
| Lasso     | Contraloría   | Upsample  | No    | 0.32        | 0.13         | 0.69         | 0.58          | 0.02   |
| Lasso     | Extension     | Raw       | No    | 0.56        | 0.33         | 0.79         | 0.68          | 0.23   |
| Lasso     | Extension     | Upsample  | No    | 0.59        | 0.39         | 0.80         | 0.71          | 0.09   |

Fig. 3. ROC curves for all models using the three outcome variables.

Additional covariates can thus lift the results by letting the model use additional information. The GBM, on the other hand, is able to explore and exploit transformations to the original data, similar to non-linear transformation and interactions in the linear model. This limits the value of additional covariates.

Results with additional performance metrics are shown in Table 5. Note that the models that use information from the Municipal Panel are indicated with the value “Yes” in column $X_c$. The results are similar to what is shown in Table 4. Indeed, inclusion in the Confecámaras list was harder to predict than any of the other two outcomes. It is telling that when we look at the 100 cases with the highest predicted probability of being listed by the Confecámaras, the precision is much higher than when looking at the top 1000 cases. In other words, between 11% and 22% of the 100 cases with the highest probability of being listed by the Confecámaras were indeed included on it, a proportion that drops to as low as 4% when we look at the top 1000 cases. This observation is confirmed by the NCDG and by the consistently low Brier scores shown in the Table. However, the precision relative to the cases with the highest 100 and 1000 predicted probabilities is much higher for our other two outcomes. The agencies would expect that, in the worst-case scenario, at least 18% of them will correspond to true potentially suspicious contracts. Finally, it is worth noting that, by comparing the rows using additional covariates from the municipal panel, we see that additional data did not significantly improve the performance of the model, and that inclusion by the Confecámaras, even with more data than before, was still hard to predict, although the precision was noticeably higher than before.
From the discussion in this section, several conclusions are worth noting. First, there may be a disagreement in the relative ranking of models using different outcomes, depending on the metric we use. However, throughout all setups, we observed performance that was sufficiently high for practical applications. Second, differences in performance across the setups were relatively small: while there are many alternatives to managing outcome imbalance, our research shows that simple approaches, such as up-weighting the minority class, work well. Moreover, we did not observe a substantial improvement in performance when using additional data. The slight impact from applying different methods of addressing imbalance was likely due to the fact that our minority class was
proportionally larger than in some of the examples frequently used in the literature, which deals with incidences much smaller than 1%. Finally, it is plausible that malfeasance and contract extensions are strongly driven by features of the contract itself (such as its cost) and not by the characteristics of the place in which the contract is executed. This would explain the weak effect of adding information from the Municipal Panel into the models. Consequently, a model such as GBM, which is able to exploit complex dependencies among the input features, is worth the additional effort in model fitting, model assessment, and model interpretation.

5.2. Model interpretation

As noted above, while lasso is easy to interpret, the GBM (our best performing model) is a black-box model that does not lend itself to an easy explanation of the predictions. However, we can still recover quantities that allow us to gain insights into what the predictive models are capturing. Moreover, burgeoning research (see Doshi-Velez & Kim, 2017; Du et al., 2018; Guidotti et al., 2019, for recent surveys) provides sophisticated discussions of methods of interpreting and explaining to humans how the GBM arrives at decisions.

We are interested here in two types of interpretations. On the one hand, and from the perspective of the final user, we are interested in knowing which variables are relevant when predicting the final outcome of a given contract. In particular, we are interested in identifying the features that are not spuriously related to the outcome and that contribute to the predictive capacity of the model. On the other hand, and for those variables that are identified as relevant, we are interested in how they affect the probability of the outcome. In other words, we are interested in understanding how much of a change in the probability of the outcome is caused by an increase or decrease of each of the inputs.

A common approach to identifying the relevant variables is to evaluate their importance, which is a model-dependent concept. In the case of the GBM, variable importance captures the number of times a variable is selected for splitting, weighted by the improvement of the model as a result of the split (Friedman et al., 2001). For lasso, importance is captured by the value of the normalized coefficient.

Fig. 5 shows the distribution of variable importance for all the models in our setup. In general, the majority of variables were dropped and unused by the prediction. Although a few more variables were used by lasso, especially when we included municipal-level information, the results in Fig. 5 indicate that a relatively small set of variables was needed to produce the performance described above, regardless of the outcome variable. Furthermore, a few variables stand out as having much more importance than the rest. Thus, it seems that a small set of factors trigger investigations of malfeasance, breach of contract, or extensions. This, in turn, creates fairly well defined types of cases.

The top variables are shown in Fig. 6. We display the top variables selected by the best model according to our results: viz., the GBM using SECON data exclusively and weighting observations to deal with imbalance. As further evidence that the method of dealing with sample imbalance does not affect the qualitative interpretation of the results, we overlay the importance of the variables selected by the models that used up-sampling. Some clear patterns can be seen. First, the variables selected by both weighting and up-sampling were remarkably similar. Of the top seven variables in the final three models at the national level, only one was not selected by the models that used up-sampling.

It is also remarkable that the budgeted cost of the contract appeared consistently in all models, with one exception (viz., the model predicting an investigation by the Contraloria using local-level data) and this variable often ranked among the top two. There is thus clear evidence that the likelihood of a troublesome outcome is related to the total size of the contract. As we show below, more expensive contracts tended to have higher risks of malfeasance, no matter what metric we used. This is an important result, as it highlights that large projects tend to have more complications and challenges. Something similar can be concluded with regard to the duration of contracts, another characteristic that ranks highly in terms of its predictive importance. Some of the most notorious corruption scandals in Colombia in recent years, such as the intervention to recover of the navigability of the Magdalena River or the construction of an important highway for public transportation in Bogota, corresponded to long and expensive projects.22

A number of variables related to the type of contract were often selected by the models. For instance, the models often picked contracts related to “transportation and storage services” (Servicios de Transporte, Almacenaje y Correo). This result is in line with the fact that in recent years, many corruption scandals corresponded to projects associated with the transportation sector.23 Contracts related to “public services for the government” (Servicios Públicos y Servicios Relacionados con el Sector Público) were also signaled by our early warning system. This is an important result, as it reveals that transactions associated with the provision of supplies as well as contracts to hire public sector staff may be malfeasant.24 This result may be a consequence of the patronage system that prevails...

22 For information on the Magdalena River scandal, see https://www.portafolio.co/negocios/empresas/le-decretan-la-muerte-a-navelena-arrastrada-por-el-caso-odebrecht-513416. For information on the Calle 26 scandal in Bogota, see https://www.elespectador.com/noticias/judicial/el-contrato-de-calle-26-se-perdieron-100-millones-de-dolares-articulo-459299.
23 Analyzing press news from 2016 to 2018, Transparencia por Colombia (2015) found that 15% of reports corresponded to the infrastructure and transportation sectors. Scandals are often associated with public works as well as with collusion between public officials and transport service providers.
24 While there are important differences between services and supply contracts, we felt it was important to include both types in our analysis. Theoretically, certain forms of corruption, such as associated with cost overruns, may become common in contracts to supply goods. In the case of service-related contracts, favoritism and nepotism appear more frequently. As our results show, both types of contracts are prone to malfeasance. Hence, an early warning system is valuable in both cases. Moreover, if corruption were more prevalent...
in developing countries like Colombia (Colonnelli et al., 2020; Gallego, Li et al., 2020), whereby jobs in the public sector are allocated to political allies or are used to extract rent.

It is noteworthy that certain variables describing the type of process used to award contracts were also frequently flagged as important predictors of malfeasance. In particular, contracts awarded through the direct selection procedure were often picked as risky. This method is the most discretionary and least competitive of all, as competition between bidders is absent under this procedure. In fact, using SECOP data, Gallego, Prem et al. (2020) showed that in the midst of the crisis caused by the COVID-19 pandemic, and after Colombian President Ivan Duque signed a decree that relaxed public procurement rules, the increase in the use of discretionary contracts was higher in more corrupt places. Similarly, Bosio et al. (2020) showed that discretion is higher and positively correlated with corruption in places with lower levels of state capacity. Moreover, Type H direct procurement, which corresponds to the hiring of natural persons for the provision of certain specific services, appeared as especially risky. Taken together, this evidence reinforces the argument regarding the connection between corruption and patronage that we discussed above (Gallego, Li et al., 2020).

Other variables related to the process were also frequently selected by the models. For instance, most models selected as a top predictor the waiting time between the date of the award of the contract (adjudicación) and the date the contract information was first published in SECOP (fecha de cargue). This delay, which we define as the “waiting period,” is an indicator of transparency and

---

25 “Contratos de prestación de servicios profesionales y de apoyo a la gestión, o para la ejecución de trabajos artísticos que solo pueden encomendarse a determinadas personas naturales. Literal H. Numeral 4 artículo 2 Ley 1150 de 2007; artículo 81 Decreto 1510 de 2013.”

26 Type C direct procurement, which corresponds to agreements and contracts between public entities, was also particularly risky according to our models.
publicity, as it signals the authorities’ diligence and timeliness when publishing contract information.\textsuperscript{27} Hence, this result directly speaks to literature on the effects of publicity requirements and public procurement (Coviello & Mariniello, 2014). Evidence suggests that more and better publicity leads to more competition and lower costs. Our models forcefully show that the waiting period is an important predictor of corruption, breach, and inefficiency.\textsuperscript{28}

\textsuperscript{27} In fact, certain municipality indicators of corruption and state capacity, such as the Open Government Index, rely on this type of measure.

\textsuperscript{28} To be clear, Fig. 6 only shows that this variable is important, but it tells us nothing about the direction of the effect. To determine this direction, it is necessary to estimate the corresponding partial dependency plot, as we do for other variables, below.

It is also significant that some of the models selected the distance to the nearest presidential election as a top predictor. This can be seen as an indication of a political–business cycle. In Colombia, regulation preventing increases in public spending before elections, the so-called ley de garantías, only applies to presidential elections, and not to regional or municipal elections.

Finally, a number of geographical predictions stand out: models often selected at least one covariate capturing specific departamentos, like Valle del Cauca (around Cali), Cundinamarca (which surrounds the capital, Bogotá), Bogotá itself, and Antioquia (which is home to the country’s second-largest city, Medellin). These are the biggest cities and departamentos in the country. At least two reasons would explain why our models assigned high values to contracts signed in populous locations. First, in these places, anti-corruption entities, like the Contraloría,
are usually more active. But this can only be part of the explanation, since some of these geographic dummies are also present in the Confecámaras and “Extension” models. The second explanation is that larger places tend to have more expensive and complex contracts, which, as we saw above, tend to have a higher risk of malfeasance. It perhaps suffices to say that corruption scandals in Colombia’s biggest cities are not uncommon.

Now that we have isolated which variables mattered the most for the predictions, we can explore the way in which the variables were related to the outcome. Thus, we created partial dependency plots, which represent the marginal value of a given variable averaging out all others (Friedman et al., 2001). Fig. 6 identifies the variables that mattered the most for the predictions, but the figure does not tell us the way in which they increased or decreased the probability of observing the outcome or by how much. For instance, according to Fig. 6, the size of the budget and the time to execute the contract are, unsurprisingly, the most relevant variables when predicting whether a contract will require an extension. What the figure does not tell us is whether this means that extensions will be required by cheaper or more expensive projects or, more interestingly, by a combination of these two variables. Figs. 7 and 8 address this question for two illustrative variables, although others could have been selected.

Fig. 7 represents for each pair of budgets and execution times the predicted probability with which each of the outcomes could happen using, in this case, a random sample of all the contracts. The figure presents the predictions by combinations of percentiles of both variables, and, interestingly enough, they represent three different types of effects. For instance, we can see that, for Confecámaras, the effect of the size of the budget monotonically increases, almost regardless of its expected duration. In other words, it seems like contracts that are likely to result in a breach are often the most expensive ones, regardless of how long they take to complete. The effect of the two variables is different in the case of malfeasance. Contracts shown in the lower-left corner and those slightly to the upper-right of the figure were more likely to be investigated: that is, short, cheap projects and long, expensive ones were more likely to be investigated by the agency. Finally, in the model predicting whether the contract will result in an extension, we observed a fairly intuitive regularity, namely that the likelihood of requesting an extension increases as both variables increase simultaneously. Thus, what the model is automatically capturing is the interaction between the two inputs: it is the combination of being both an expensive and a long-term project that increases the probability of requesting additional resources—more so than being either an expensive short-term project or an inexpensive long-term project.

Finally, Fig. 8 shows the partial effect corresponding to each departamento of Colombia, that is, the risk of a troublesome outcome for each departamento. It is interesting to observe that there is a very small differential effect by departamento in the case of Confecámaras, with the exception of Bogotá and its surrounding region (as indicated already in Fig. 6). However, extensions to contracts and investigations of malfeasance show a clearer group of higher-risk departamentos, such as Antioquia, Bogotá, and Cundinamarca. Although each of the departamentos has a different number and distribution of the type of contract, Fig. 8 can be seen as a first approach to a territorial risk score to be used by central government agencies when allocating resources to fight malfeasance and inefficiency.

6. Conclusions

Transparency is crucial to curb undesirable outcomes in public procurement, such as malfeasance, breaches of contract, and general inefficiency. Web-based platforms to register and report public transactions have become popular in both developed and developing countries, as they enable anti-corruption agencies, watchdogs organizations, and citizens to monitor closely the process and the decisions through which governments acquire goods and services. The combination of more and better information, increased computational capabilities, and more flexible statistical techniques to analyze administrative data represents a unique opportunity to fight corruption using tools of the so-called big data revolution.

We developed early warning models for this type of analysis using a unique dataset of over two million public procurement contracts in Colombia from 2011 to 2015. We trained two representative machine learning models to identify as early as possible those contracts that are likely result in undesirable outcomes for the government. We believe that this type of analysis is useful for public management for two reasons. First, predictions with our models represent risk scores that can be used by authorities when selecting which contracts they should monitor and audit. Our outcome variables approximated different aspects of malfeasance, from the most to the least serious cases: proven corruption (using data from the Contraloría General de la República), breach of contract (provided by the Confederación Colombiana de Cámaras de Comercio), and inefficiency (which we approximated through extensions to the originally stipulated budget or to the duration of the contract). While these types of outcomes can show empirical correlations, they measure concepts that call for different actions by government authorities: contracts that are predicted to result in corruption may trigger visits by auditors, whereas if they are predicted to request extensions, an alert to the contracting agency may be sufficient.

Second, the methods we used allow us to describe which variables—and in which way these variables—contribute to the likelihood that a contract will be problematic, which is very useful from the perspective of policymakers. This information can guide discussions on reforms to curb corruption. In our case, for instance, variables associated with projects such as their size or duration were important predictors of malfeasance. Also, the time lag between adjudicating the contract and the nearest election showed high predictive value. Naturally, these models are not causal; their main goal is prediction. However, combining these methods with traditional causal inference techniques may shed some light on how to design policies to limit malfeasance. In the presence of
hundreds or thousands of features, variable importance, in the way we used it here, suggests a way to start. Practitioners should differentiate between features that can potentially cause an outcome, features that may be a consequence of a common cause of the outcome, and features that are themselves consequences of the outcome (Guyon et al., 2007). Of course, only the first set of variables are candidates for policy intervention. In other words, we do not see our methods as substitutes. Instead, we believe that they should be used to complement other methods, to improve the policymakers’ toolkit.

This discussion does not imply that corruption can be completely thwarted through the use of public procurement data and machine learning methods to predict problematic contracts. The methodology still faces many challenges. First, self-selection by the government entities that register information with these platforms is a major challenge. Second, outcome variables in this domain are often difficult to obtain. Third, authorities need to prioritize and balance precision and recall and weigh the costs and benefits of using aggressive or passive classifiers. Finally, it is important to remember that contractors and malevolent public servants are not static agents. They can anticipate which traits are more likely to trigger an alarm by an early warning system. Thus, for the models to remain useful, they should be periodically adapted to new conditions and behaviors. At the end of the day, anticipating which public procurement contracts are at risk is essential in order to improve the quality of governance and public service delivery.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2020.06.006.

References

Acevedo, K. M., & Bornacelly, I. D. (2014). Panel municipal del CEDE. Technical report, Universidad de los Andes-CEDE.

Ades, A., & Tella, R. D. (1999). Rents, competition and corruption. American Economic Review, 89(4), 982–993.

Adsera, A., Boix, C., & Payne, M. (2003). Are you being served? Political accountability and quality of government. Journal of Law, Economics, and Organization, 19(2), 445–490.

Anderson, T. (2009). E-governments as an anti-corruption strategy. Information Economics and Policy, 21(3), 201–210.
Henao, J. C., & Isaza, C. (2018). **Corrupción en Colombia.** Universidad Externado de Colombia.

Ivanov, K. (2007). The limits of a global campaign against corruption. In S. Bracking (Ed.), **Corruption and development: The anti-corruption campaigns.** Palgrave Macmillan.

Johnston, M. (2001). Syndromes of corruption: Wealth, power, and democracy. Cambridge University Press.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. **Quarterly Journal of Economics,** 133(1).

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. **American Economic Review: Papers and Proceedings**, 105(5), 491–495.

Kotsiantis, S. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. **Artificial Intelligence Review,** 37(4), 331–344.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. **Proceedings of the fourteenth international conference on machine learning.**

Kuhn, M. (2018). **caret:** Classification and regression training. R package version 6.0-80. [https://CRAN.R-project.org/package=caret](https://CRAN.R-project.org/package=caret).

Kuhn, M., & Johnson, K. (2013). **Applied predictive modeling.** Springer.

Lagunes, P. (2017). **Guardians of accountability: A field experiment on corruption and inefficiency in local public works.** Working paper.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In **ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).**

Lawson, L. (2009). The politics of anti-corruption reform in Africa. **Journal of Modern African Studies,** 47(1), 73–100.

Lee, K. (2009). A final flowering of the developmental state : the IT policy exeperiment of the korean information infrastructure, 1995-2005. **Government and Information Quarterly,** 26, 567–583.

Lewis-Faupel, S., Neggers, Y., Olken, B., & Pande, R. (2016). Can electronic procurement improve infrastructure provision? Evidence from public works in India and Indonesia. **American Economic Journal: Economic Policy,** 8(3), 258–283.

Lichard, G., & Fernades, G. (2019). *The dark side of the contract: Do government audits reduce corruption in the presence of displacement by vendors?* Working paper.

Ling, C. X., & Li, C. (1998). Data mining for direct marketing: Problems and solutions. In **KDD: Vol. 98** (pp. 73–79).

Lopez-Iurrugi, F., & Sanz, I. (2018). Predicting public corruption with neural networks: An analysis of spanish provinces. **Social Indicators Research,** 140(3), 975–998.

Mauro, P. (1995). Corruption and growth. **Quarterly Journal of Economics,** 110(3), 681–712.

Mena, J. (2011). **Machine learning fornsics for law enforcement, security, and intelligence.** Taylor and Francis Group.

Mironov, M., & Zhuravskaya, E. (2016). Corruption in procurement and the political cycle in tunneling: evidence from financial transactions data. **American Economic Journal: Economic Policy,** 8(2), 287–321.

Molnar, C. (2019). **Interpretable machine learning.** https://christophm.github.io/interpretable-ml-book/.

Muchinski, D., Siroki, D., He, J., & Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. **Political Analysis,** 24(1), 87–103.

Mungiu-Pippidi, A. (2015). *The quest for good governance.** Cambridge University Press.

Myerson, R. (1993). Effectiveness of electoral systems for reducing government corruption—A game-theoretic analysis. **Games and Economic Behavior,** 5(1), 118–132.

Olken, B. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. **Journal of Political Economy,** 115(2), 200–249.

Pathak, R. (2009). E-governance to cut corruption in public service delivery: A case study of Fiji. **International Journal of Public Communication and Technology (ICT) and social capital.** **International Review of Administrative Sciences,** 75, 99–116.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B. Statistical Methodology,** 67, 253–273.

We, S. (2000). Natural openness and good government. Working paper, NBER.