Preventing rather than Punishing: An Early Warning Model of Malfeasance in Public Procurement

Jorge Gallego
Universidad del Rosario

Gonzalo Rivero
Westat

Juan David Martínez
Alkanza

September 19, 2018

Abstract

Is it possible to predict corruption and public inefficiency in public procurement? With the proliferation of e-procurement in the public sector, anti-corruption agencies and watchdog organizations in many countries currently have access to powerful sources of information. These may help anticipate which transactions become faulty and why. In this paper, we discuss the promises and challenges of using machine learning models to predict inefficiency and corruption in public procurement, both from the perspective of researchers and practitioners. We exemplify this procedure using a unique dataset characterizing more than 2 million public contracts in Colombia, and training machine learning models to predict which of them face corruption investigations or implementation inefficiencies. We use different techniques to handle the problem of class imbalance typical of these applications, report the high accuracy of our models, simulate the trade-off between precision and recall in this context, and determine which features contribute the most to the prediction of malfeasance within contracts. Our approach is useful for governments interested in exploiting large administrative datasets to improve provision of public goods and highlights some of the tradeoffs and challenges that they might face throughout this process.

1. Introduction

"Division of Structured Operations" was the name of the department created by the Brazilian construction company Odebrecht to deal with politicians in several countries where the company based its operation. The recent corruption scandal, that links this company to several governments in Latin America, reveals that this division actually was the department of bribery of the corporation. As it was confessed by former CEO Marcelo Odebrecht, the firm paid millions of dollars to politicians in countries like Brazil, Colombia, Peru, Venezuela, Panama, among others, in exchange for favorable conditions in public procurement processes. In Peru, for example, former president Pedro Pablo Kuczynski resigned after heated controversy surrounding his relationship with Odebrecht, while other former Peruvian presidents Ollanta Humala and Alejandro Toledo also face serious judicial accusations for similar reasons. This example, as well as many other corruption scandals throughout the world, underscore the importance of curbing malfeasance in public procurement in order to reach development and well-being objectives (Mauro, 1995; Rose-Ackerman, 1999; Bardhan, 2006; DiRienzo et al., 2007).

Several measures have been suggested to curb corruption, with transparency of rules, laws, and transactions appearing as an important element of the equation (Ades and Tella, 1999; Treisman, 2000; Wei, 2000; World Bank, 2013). The consolidation of new sources of data and information—boosted by recent technological improvements—and of new and improved methodologies to analyze them, represent an opportunity for governments to enhance its fight against corruption and inefficiency (West, 2004; Anderson, 2009; Berton et al., 2010; Prasad and Shivarajan, 2015). Machine learning techniques applied to novel datasets by both practitioners and academics, have proved effective in several areas of public service delivery such as security (Mena, 2011), education (Kotsiantis, 2012), health (Kleinberg et al., 2015), conflict (Muchlinski et al., 2016) and justice (Kleinberg et al., 2018). Until recent years, quantitative policy research has mainly focused on causal inference, with program evaluation—determining if policy X has the expected impact on outcome Y—prevailing in many settings. But in many other instances, prediction may be more important than causality (Kleinberg et al., 2015). Anticipating with precision the realization of an outcome variable is crucial for governments that need to allocate efficiently scarce resources in order to maximize well-being. Prevention, detection, and sanctioning of inefficiency, fraud, and malfeasance, constitute a perfect example of an area where prediction is essential to optimize the provision of public goods.

In recent years, several countries have consolidated web-based platforms for public procurement, in which public agencies have the obligation of registering all their contracts and economic transactions. In some places, this information is open source, meaning that anti-corruption agencies, civic society groups, watchdog organizations, and citizens in general, have the chance of scrutinizing these transactions. Nonetheless, the volume and complex-

ity of this information requires special skills, techniques, and computational capabilities in order to analyze this information. Machine learning techniques, through algorithms such as decision trees, random forests, regularization methods, among many others, seem to be particularly useful for this purpose. Using data from these e-procurement platforms, such algorithms may be used to train predictive models of malfeasance, as a function of the different characteristics that define these transactions. So far, some attempts have been conducted in this direction, for example at the level of international development contracts (Grace et al., 2016) or provinces within a country (Lopez-Iturriaga and Sanz, 2017).

We take this analysis one step further, by implementing machine learning techniques to predict inefficiency and malfeasance, at the contract level, using information from Colombia's e-procurement system, the *Sistema Electronico de Contratación Pública* (SECOP). We consolidate a unique dataset of more than two million public contracts between 2011 and 2015, and merge this information with judicial-level data of contractors and aggregate characteristics of the municipalities in which these transactions take place. Because contracts with proved judicial problems constitute a minority of total observations, we must handle the methodological challenge of class imbalance typical of this type of applications. As a consequence, we are able to estimate models that predict whether a contract becomes inefficient[1] or offered to contractors that face judicial investigations, according to anti-corruption agencies. We claim that training models in this way is useful for at least two reasons from a public management perspective. First, the adjusted values of these models represent *risk scores*, that may guide anti-corruption agencies and watchdog organizations in the selecting contracts for monitoring and audit purposes, based on what they have found in the past. Second, our models include an analysis of feature importance—i.e. which variables associated with contracts are more important to predict corruption investigations and inefficiency. Determining which variables correlate the most with corruption is important, as it can guide discussions on the the type of institutional reforms that are needed to curb malfeasance.

Our models have high levels of performance, exhibiting levels of the *Area Under the Curve* (AUC), a common measure used in these settings, always above 80%—and in some cases above 90%—when predicting inefficiency or judicial investigations. Moreover, our framework proves useful to balance the tradeoff between precision and recall—type I and type II errors—which in this context simplify into having more aggressive or more passive classifiers. In many cases governments may prefer more aggressive classifiers that minimize the expected number of false negatives, that in this case represent malfeasant contracts that are not forecasted as such. However, if resources for monitoring and audits are scarce, it might be preferable to minimize false positives. We present simulations in which we vary

[1] Inefficiency is measured as contracts that require more money or more budget than what was originally stipulated. See a deeper discussion in the next section.

the relative cost of false positives, that shed some light on the mechanisms behind this tradeoff.

Despite the large number of features that we have for Colombian contracts (more than 300), our models show that only a limited number of characteristics matter when predicting inefficiency and corruption detection. Typical and important characteristics of contracts, such as their size, the lag between the day the contract was awarded and the first day of implementation, the distance to the nearest election, or geographical and sector-specific patterns, are some of the key features that matter the most when predicting sanctions and inefficiency. Hence, a combination of the main traits of these projects, and certain political characteristics associated to electoral cycles and the seasonality of procurement, determine to a great extent the result of these public transactions.

This paper is composed of seven sections including this introduction. In section 2, we describe some of the theoretical foundations that relate to our approach and discuss some challenges that practitioners will face when implementing machine learning algorithms to predict corruption detection and inefficiency. Section 3 characterizes the background surrounding the Colombian case, while section 4 describes the data employed in this study, which represents a combination of public procurement information, judicial evidence, and aggregate municipality covariates. Section 5 describes the models implemented in the analysis, including the techniques used to mitigate the issue of class-imbalance and the machine learning algorithms that we trained. We present the main results of our analysis in section 6, reporting the overall performance of the models, characterizing some of the main tradeoffs that policy-makers will face when using these tools, and discussing which variables are more important at predicting our outcomes. We conclude in section 7.

2. Theory

Corruption, the use of public office for public gain (Rose-Ackerman, 1999), is a pervasive problem prevalent in the developing world. Inefficiency is a phenomenon closely associated with corruption, and in many cases, the former is a manifestation of the latter. In fact, for some authors (Lagunes, 2017) corruption can be classified as *active waste*, in that public servants deliberately interfere in the process of public service delivery for personal benefit. In contrast, inefficiency is equivalent to *passive waste*, whereby distortions are not the result of actions intently aiming to increase personal gains. In any case, there is evidence supporting the existence of a negative correlation between corruption and inefficiency (Dal Bo and Rossi, 2007).

As stated by Besley (2006), corruption and inefficiency are common in countries where governments are not held accountable. Consequently, both phenomena have important negative effects on the economy. There is a negative correlation between corruption and

growth (Mauro, 1995), the quality of democracy and the rule of law (Mungiu-Pippidi, 2015), and service delivery in terms of security (Condra et al., 2016), education (Dufflo et al., 2012), and health (Chaudhury et al., 2006), among other dimensions. Unsurprisingly, the puzzle of what drives "good governance" is crucial to understand why some countries are more developed than others, although, until recently, scholars have paid more attention to the internal workings of the state and the individuals in charge of providing public services (Finan et al., 2016).

If good governance is crucial for development, and yet corruption and inefficiency abound, a major task for governments and practitioners is to find ways to minimize both phenomena. The problem has been structured as a principal-agent relationship, in which information asymmetries explain the prevalence of inefficient outcomes (Riley, 1998; Andvig and Fjeldstad, 2001; Johnston, 2001; Ivanov, 2007; Lawson, 2009). Either if we consider the relation between governments and voters (Myerson, 1993; Adsera et al., 2003; Besley, 2006), or between elected officials and appointed bureaucrats (Becker and Stigler, 1974; van Rijckeghem and Weder, 2001), lower levels of information and lack of transparency translate into poor delivery of social services. It has been accepted in the literature that in order to discipline bureaucrats it is necessary to achieve a balanced combination of incentives, monitoring, and sanctioning.

Nonetheless, scholars have debated on whether top-down accountability mechanisms are more or less efficient than bottom-up strategies. As revealed by his seminal work on public works in Indonesia, Olken (2007) claims that corruption and inefficiency are lower when top-down audits are used to supervise projects, in contrast to grassroots accountability strategies. Hence, this literature underscores the importance of government agencies and watchdog organizations that aim to gather information on the performance of public servants. Other studies, such as Ferraz and Finan (2008), Bjorkman and Svensson (2009) and Chong et al. (2015), show how information is crucial to curb corruption, as better informed voters are able to hold accountable elected officials more effectively. Other authors show how the cooperation between civil society organizations and government agencies is capable of reducing corruption and inefficiency in developing countries (Lagunes, 2017).

Consequently, no matter if it is through top-down or bottom-up strategies, information and transparency are key elements associated with good governance (Prasad and Shivarajan, 2015). In this context, the consolidation of new technologies and sources of information increase the opportunities for holding bureaucrats accountable. ICTs and e-governments provide new tools to fight against malfeasance. In fact, cross-country evidence suggests that there is negative correlation between corruption and the free flow of information. Countries that embrace transparency tend to produce more information and are more likely to use it and share it with the general public (Lord, 2006; Berton et al., 2010). In recent years, web-based platforms have been established throughout the world, in order to keep track of procurement and economic transactions carried out by governments. One of the long-term

goals of these platforms is to reduce the information asymmetries that exist between officials and the public. E-government has proved to have the potential of reducing corruption by enhancing efficiency and transparency in various contexts (Berton et al., 2010), such as India (Bhatnagar, 2003; World Bank, 2004), Pakistan (Anderson, 2009), Chile (Shim and Eom, 2009), Fiji (Pathak, 2009), and Korea (Lee, 2009), just to mention a few.

Technological change, and in particular the creation and development of the internet, has enabled governments to create e-procurement systems and platforms.[2] As a result, new and powerful databases, similar in nature to SECOP, have emerged all around the world. This has created new opportunities for monitoring and accountability, because in many countries, including Colombia, both anti-corruption agencies, as well as the civil society, have access to this information. But new challenges arise as well, such as what has been dubbed as big data. As governmental transactions are registered and recorded, millions of bytes of information are being produced every minute. Storage, analysis, and interpretation of this information are not necessarily straightforward tasks.

In this context, machine learning techniques provide useful tools for prediction and anomaly detection. Prediction has proved to be effective in other areas of government performance, such as security (Mena, 2011), justice (Kleinberg et al., 2018; Berk, 2012), health (Kleinberg et al., 2015), education (Kotsiantis, 2012), peace (Gallego et al., 2017), among other fields. In all these cases, policy makers are able to anticipate patterns and to predict outcomes, which serve as a guidance to implement welfare-enhancing policies. For instance, governments anticipating in which streets urban crime is more likely to be committed, are able to deploy police units beforehand to such vicinities. Similarly, if school officials can predict which students are more likely to drop out, special targeting of these children may prevent from losing them. Similar applications can be rolled out in terms of transparency and good governance.

As we show in this paper, machine learning techniques can be useful when fighting against corruption, for at least two reasons: risk detection and institutional reform. First, keeping institutional rules constant, predictive models as the ones we employ in this paper, are valuable in the short and middle terms because they enable the creation of corruption and inefficiency risk scores. This works in a similar fashion as credit scores in finance: financial institutions estimate the probability of default of a client thanks to information collected at the individual level. In fact, risk scores of this nature have been used elsewhere at aggregate levels, as in the case of Spanish provinces (Lopez-Iturriaga and Sanz, 2017).

The originality of our approach is that we calculate the scores at the contract level. This represents a more granular estimation that is useful because many subnational units or governmental agencies are more likely to commit fraud in certain areas but not in others. A

[2]Social media represents an additional web-based contribution to the fight against corruption. Through blogs, wikis, social networking sites, micro-blogging services, and multimedia-sharing services, millions of users throughout the world have found new spaces to hold their governments accountable.

malfeasant government may find it easier to extract resources from an infrastructure project rather than from other types of investments. Clearly, determining which contracts are more likely to be involved in fraudulent activities is useful for anti-corruption agencies and watchdog organizations, because audits are a scarce resource and not all contracts can be monitored and investigated. Additionally, early-warning messages can be sent beforehand to officials in charge of those contracts that are classified as risky, in order to deter corrupt behavior on their side. Hence, these tools are useful both to deter and to detect malfeasance in public procurement.

Second, predictive models are useful in the fight against corruption because they can guide institutional reform in order to curb malfeasance in the medium and long terms. An interesting feature of several machine learning algorithms is that they identify the correlates of the outcome variables. In other words, through *variable importance* indicators, these models help researchers determine the level of predictive contribution of each covariate. In our case, this information is useful because it indicates which contract-level or municipal-level characteristics are more strongly associated to the occurrence of corruption and inefficiency. As such, our results could inspire governments seeking to carry out institutional reforms in order to reduce corruption and inefficiencies in public procurement.

However, it is important to acknowledge that the use of predictive models as a tool against malfeasance is not exempt of challenges and pitfalls. First, there is a problem of self-selection in the provision of information to platforms like SECOP. In the Colombian case, subnational administrative entities and other agencies are in charge of filling out the information required by the system. Naturally, there is some room for strategic misinformation in some cases. Of course, information can be compared using other sources, but in a context of millions of contracts this can prove to be a difficult and time-consuming task. Moreover, incompleteness is another possible strategy, and one would think that officials more deeply compromised in corruption and inefficiency have the strongest incentives to hide information. In fact, during the early stages of SECOP, a large fraction of agencies did not report on time (or at all) their information. In such cases, it is crucial that central governments enforce the mandate of providing on-time truthful information on public procurement.

A second challenge concerns the availability of outcome variables. In the Colombian case for instance, SECOP provides information on inefficiency, as it is possible to determine which contracts present delays or end up costing more than what was originally planned. However, direct measures of malfeasance, such as contracts that end up being investigated by anti-corruption agencies, are not provided by these platforms. Hence, it is crucial to match information from different sources, in order to obtain the data needed to feed in the models. However, in many developing countries, information from the judicial system at the contract level is not centralized nor organized, or simply is difficult to obtain. Consequently, cooperation between different government offices is crucial to consolidate high quality data.

The third challenge is related to a common trade-off in machine learning applications: the distinction between *precision* and *recall*, or, in other terms, *false positives* and *false negatives*. In the context of anti-corruption models, a very "aggressive" classifier will classify many contracts as potentially corrupt and inefficient. In such cases, the rate of false positives may be high. If governments are not highly budget-constrained, it might be preferable to have this type of model, in order to reduce the number of malfeasant contracts that are not detected as such. This can be desirable when the goal is to audit a large number of contracts, in order to signal the strength of the agency and deter corrupt behavior.

Conversely, a very conservative model will minimize the number of false negatives, i.e. the amount of contracts that are incorrectly classified as potentially malfeasant. When audits are a scarce resource for governments, it may be more efficient to audit contracts signaled as malfeasant by the model with a high probability. In this way, less resources will be wasted on investigating contracts classified as malfeasant but which are likely not. In the end, depending on the objective function of anti-corruption agencies, the parameters of the models can be fine-tuned in order to maximize the level of sensitivity or specificity. Naturally, the challenge is to establish beforehand such an objective function.

A fourth and final challenge needs to be discussed: contractors and bureaucrats are not static agents that do not adapt to new conditions. One would expect that if these algorithms start to be used by agencies, those actors involved in corruption may adjust their behavior in order to reduce the probability of being detected by the model. Consequently, the application of these machine learning models needs to be adaptive and dynamic as well. Training and test datasets need to be fed constantly with new information, in order to account for changes in behavior on the other side of the relationship. Moreover, information sources like SECOP are not static either, as new variables are incorporated once these systems consolidate. For instance, the new version of the Colombian platform, SECOP II, currently includes information on the number of bidders in public auctions. This variable is very important in order to predict corruption, as in many cases those contracts in which malfeasance takes place are characterized by just having one bidder. In the end, predictive models need to adjust to changes in behavior and sources of information.

3. Background

Sistema Electrónico para la Contratación Pública (SECOP) is the name of the first attempt by the Colombian government to digitalize and enhance the monitoring of public procurement in the country. SECOP relies on two fundamental pillars: (i) the publication of contracts made between the state and private contractors; and (ii) notices and information regarding awards of these contracts. The mission of SECOP is to shed light on the partnerships and contracts that public entities engage through the detailed exposure of

the awarding process. As a mandatory step in the awarding process, government agencies must record, store, and publish all procurement actions, documents, and changes using the SECOP platform. This may include offers, contracts, evaluation reports, calls for tenders, studies and other documents related to the contracts, among other documents and actions that took part during the process. As a consequence of the implementation of the SECOP platform, the public availability of procurement information in Colombia grew by 286% between 2011 and 2014 (OECD, 2016).

Public acquisition laws in Colombia allow government agencies to use several types of procurement procedures, given 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. 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. The abbreviated choice method is intended to be used to contract standardized products or services. Merit selection, on the other hand, is a tool to hire consultants and advisors for public entities. Direct selection is only allowed in cases of manifest urgency, lending contracts, or inter-administrative contracts. Finally, the minimum-value contract method is used in cases where the procurement is below a specified amount. In theory, the common practice should be that the selection of contractors should go through the public tender processes unless the context allows something different. In practice, government agencies find ways to use other procedures that may cover malfeasance, such as the direct selection method.

In Colombia, several agencies have a mandate to prevent and punish inefficiency and corruption. The Office of the Controller General (CGR) is in charge of fiscal control in the country, and, as such, its main task is to seek for the proper allocation of public funds and resources. Consequently, this office audits government entities, which may result in warnings and further investigations. These, in turn, may lead to criminal prosecutions by the Office of the Attorney General against public servants or private contractors. Additionally, the process of fiscal control is also implemented in a decentralized a way. Departamentos[3] and major municipalities, for instance, have their own *Contralorías*, with the same task of preventing and punishing corruption, but their mandate is limited to a specific territorial entity.

[3]Departamentos are the equivalent to States in the U.S.

4. Data

In the analysis that follows, we use data on public procurement in Colombia between 2011 and 2015 coming from the SECOP database. The SECOP system was rolled out in two different phases. SECOP I served as a tool to collect and publish the entire contractual activities of all government entities, from planning stages to liquidation. The database was designed with respect to two main goals: "tidiness in the selection of vendors and more advantageous conditions for the government"[4] in an effort to increase transparency, curtail inefficiency and, consequently, reduce corruption in public procurement. As part of its modernization goals, SECOP II was developed to offer a single point through which vendors and government contractors can access requests from the government, submit their proposals, and monitor their progress. As a result, the SECOP database compiles every single transaction between the Colombian government and its suppliers regardless of the value or the transaction or the type of good or service acquired. The database included in 2011 has a total of 195,135 contracts with a total value of COP $33,093 billion. In 2015, the last year in our database, it totaled 886,242 contracts representing an aggregate value of COP $121.255 billions.

The information contained in the SECOP database is publicly accessible through an online platform and monthly copies of the data are posted to the open data platform of the Colombian government. (Datos Abiertos de Colombia). In the analysis that follows, we focus on the data between 2011 to 2015 which contains 58 variables and 2,241,271 observations, each of them corresponding to a transaction between a government entity and a supplier of a service or a good.[5]

The data provides information on the government entity that purchases the service or good, whether it is a national or a subnational entity (for instance, a central or a local government), and basic administrative information about the contractor. All purchases are identified by the contract's approval date, the date on which the execution of the contract begins, the duration of the contract that is specified in the requisition process, and the date on which the contract was considered to be fully executed. For each contract we also know the type of process that was used in the procurement and the current status of the process (e.g. convened, adjudicated, liquidated, etc.), a UNSPSC identifier for the good or service under contract (along with a short text description of the objective of the contract), the origin of the resources used to carry out the purchasing process, the planned budget, and the value that was finally awarded in the contract.

[4] Website of the Ministerio de Comercio, Industria y Turismo de Colombia

[5] We restrict our analysis to up to 2015 for two reasons, despite the fact that more recent years are available in the SECOP database. First, investigations and prosecutions from anti-corruption agencies present some natural time lags. Second, in recent years SECOP II has become more popular, so, for comparability reasons, we focus on contracts using SECOP I.

Besides using this contract-level information provided by SECOP, we use a battery of municipality-level characteristics coming for the Municipal Panel from Universidad de los Andes. This dataset provides a comprehensive number of indicators across different dimensions, including socio-demographic, economic, fiscal, health, education, and electoral characteristics, among some others. The analysis being conducted at the contract level, we assign to each contract the characteristics of the municipality in which it was executed. Hence, when including municipality-level predictors, we restrict the analysis to contracts executed by regional or local governments, excluding central-level state agencies. We end up with a dataset that contains 188 predictors if municipality-level characteristics are not included, and 320 predictors if we include these.

Concerning outcome variables, we use three different indicators of corruption and/or inefficiency. First, SECOP provides information on additions to the contracts in either time or money. Extensions to contracts are considered a sign of inefficiency and they are potentially linked to irregularities of the contract. In addition, we use two other variables closely linked to corruption. First, using data from audits carried out by the Office of the Controller General of Colombia and the associated regional offices, we have access to the list of contractors that have been sanctioned by these entities. Note that in this case the information is available at the contractor level—and not at the contract one—meaning that our outcome variable indicates whether a contract was signed with a contractor (natural or legal person) that later was sanctioned for irregularities in the execution of public resources.[6]

Second, we build a second indicator of corruption, using data from the *Confederación Colombiana de Cámaras de Comercio* (Confecámaras), a board of trade in the country. Confecámaras compiles information of contractors that have received fines and sanctions from different government agencies—including but not restricted to the Controller's Office—in order to prevent territorial agencies from contracting these firms or persons. This information is provided by public agencies themselves to Confecámaras, so unlike the information from the Office of the Controller General, it has a decentralized, and perhaps, broader flavor. Here again, the outcome corresponds to contracts that were signed with a contractor appearing in the "black list" of Confecámaras.

5. Models

We predicted the likelihood of each of the three outcomes discussed in the previous section using three popular machine learning models. In particular, we used a lasso logistic

[6]Naturally, it would be ideal to have more fine-grained information on corruption outcomes at the contract level, as contractors may have received sanctions for a particular contract, and none for other contracts. However, information at such a level is currently unavailable.

regression, a conditional inference tree, and a gradient boosting machine (GBM). Each of the models present different advantages for practical purposes.

Lasso is a natural generalization of the popular logistic regression that includes penalization component on the total size of the coefficients that helps with variable selection by pushing some terms to zero.[7] By doing that, it prevents the model from being more complex than needed (Tibshirani, 1996) (Tibshirani, 1996). Such a simple strategy, which has a natural connection to Bayesian statistics, has been proven very successful in domains in which the number of covariates may be larger than the number of observations. In addition to its close relation to models common in the statistical toolkit of social scientists, the lasso is very fast to fit even to a large number of observations and it is relatively easy to interpret.

We also used a classification tree model, a very common approach that tries to split the data by using cut points for each variable in order to maximize predictive accuracy. At each step of the process, the model tries to find the best variable to split the data into observations that are most different relative to the outcome variable. The resulting structure is a non-parametric structure with a series of branches representing the optimal decisions estimated by the tree that result in leaves that assign labels to groups of observations. Classification trees are well-studied approaches, relatively easy to fit, and above all, offer a very simple interpretation of the resulting model even for people with no statistical training. We chose a particular flavor of classification trees called *Conditional Inference Trees*[8] which performs splits based on significance testing. Among other advantages, it avoids variable selection bias that is induced by older approaches that tend to overselect categorical variables with many categories (Hothorn et al., 2006), which are common in datasets.

The final model we tried is a *Gradient Boosting Machine* (GBM), a very popular model in industry applications. The model consists of a potentially large number (from a few dozens to several thousands) of trees with very few splits grown sequentially so that each tree's goal is to fit the residual of previous steps. The GBM belongs to a family of models that combine "weak" learners—in this case, shallow trees—into a "strong" learner. The GBM achieves low generalization error though a number of regularization techniques that avoid that the combination of trees overfit the training data[9].

We fit our models to a random sample of 200,000 observations, and we set aside 25% of the observations as validation set (see below).[10] The remaining 150,000 observations were used

[7] We used the implementation in the `glmnet` package (Friedman et al., 2010).

[8] We used the implementation in the `party` package (Hothorn et al., 2006).

[9] We chose the traditional `gbm` (Ridgeway, 2017) instead of the more popular `xgboost` (Chen 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 interpretation stage, the user must reconstruct the original categorical variable, even if some categories are not included in the final model.

[10] Models were fit using 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

for model fitting using 5-fold cross validation, in which we tested a grid of model-specific parameters that controlled the complexity of the model.[11]

The optimal combination of parameters was picked using the "one standard deviation rule" (Friedman et al., 2001a) . In other words, we did not select the combination on regularization parameters that produced the best overall fit but instead one combination at random among those close to the optimum. This approach protects us further against mimicking characteristics specific to our sample of observations (Cawley and Talbot, 2010). All the performance measures reported in the rest of the paper come from the holdout sample and, in consequence, should approximate well the performance of the models.

A further note about the data is needed. As indicated above, administrative investigation of procurement malfeasance, as well as the extension of the term of the original contract, are rare events. In consequence, all of our dependent variables are heavily biased in favor of the "negative outcome," i.e., normal cases, as can be seen in Table 1. More than 95% of the contracts in SECOP are not listed by Confecámaras or the CGR and around 11% of the contracts receive an extension. This unbalance is potentially problematic for classification tasks (Kuhn and Johnson, 2013), as models can attain high by predicting all instances as belonging to the majority class, disregarding the minority class entirely. A number of remedies have been suggested in the literature.

Table 1: Number of observations in each of the outcomes

	Confecámaras	Contraloría	Extension
Negatives	2,217,692	2,202,513	1,989,784
Positives	23,579	38,758	251,487

One family of solutions consist on *resampling* observations from the training dataset in order to achieve a more balanced distribution of the outcome variable (Van Hulse et al., 2007). This can be achieved by *down-sampling* the majority class (Kubat, Matwin, et al., Kubat et al.) or *up-sampling* the minority class (Ling and Li, 1998). The SMOTE algorithm (Chawla et al., 2002) that we use below combines both approaches through the creation of *synthetic* cases in the neighborhood of the other observations. In order to prevent data leakage from the training to the test samples, the SMOTE algorithm is not used as a pre-processing step, but instead as part of the cross-validation procedure.

The second solution we used consisted in applying differential weight costs for each of the two outcome values to ensure that the classification model does try to attain high performance by simply increasing the Type II error (cases with extensions or being investigated by corruption that are predicted with a negative outcome). We used a weight sufficient to

[11] For instance, the weight of the penalization factor in the lasso model, the significance level required for making an additional split in the conditional inference tree, or the maximum depth of each tree in the GBM, among others.

increase the weighted proportion of the corruption/extension cases to 25%. In consequence, we used a weight of 25 for the variables collected by the CGR and *Confecámaras* and a weight of 10 for the variable that captures an extension to the contract.

In consequence, for each of the three dependent variables, we trained two different setups depending on how we dealt with the unbalance in the outcome variable using two separate approaches. In one case, we trained the model using the original raw data with cost-sensitive classification. In the other, we processed the data using the SMOTE algorithm.

6. Results

In this section, we present the predictive models described above. We first discuss model performance and the decisions that we took to select our final models, as well as an evaluation of the implications of assuming different cost structures. In Section 6.1.1 we discuss the models that we fit to a random sample of all the contracts in the SECOP database. In Section 6.1.2 we focus on a random subsample of contracts at the local (*municipio*) level. An interpretation of the results is presented in Section 6.2.

6.1. Model fit

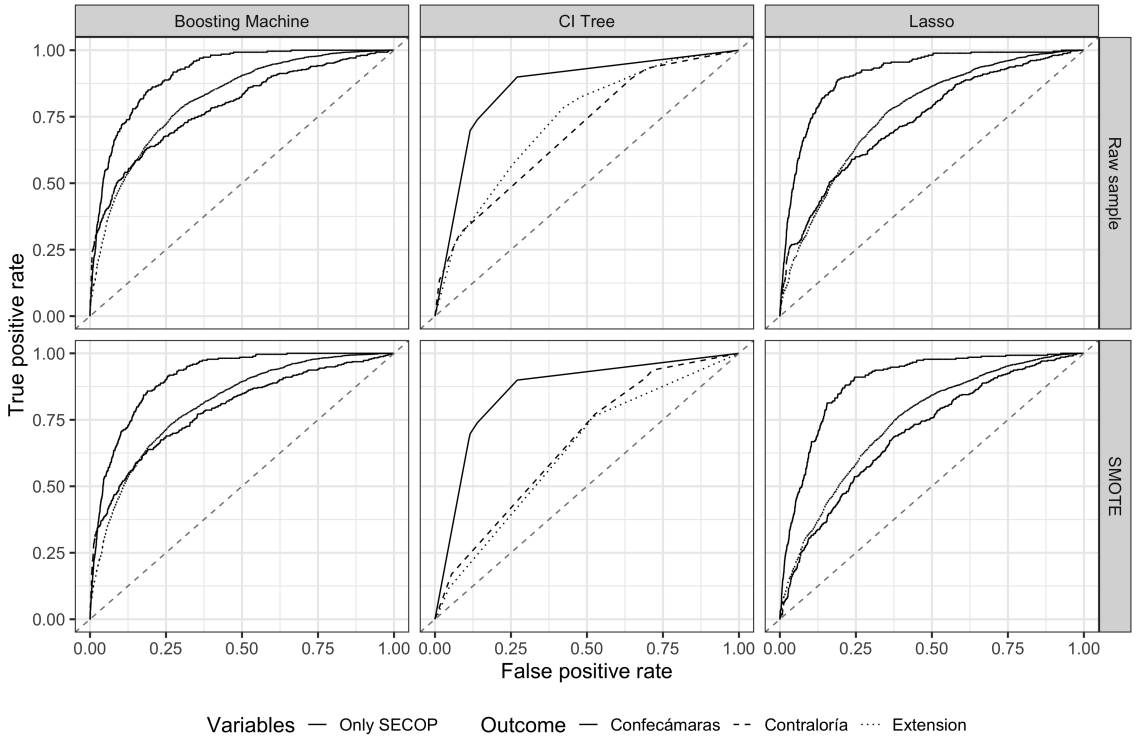
6.1.1 All cases

In Figure 1 we show the ROC curves for each of the models calculated using data in the holdout set. The ROC is a common measure of performance for binary classification models. It represents, for all possible probability cutoff points, the true positive rate (the proportion of actual detected corruption cases that are correctly classified) and the false positive rate (the proportion of incorrectly guessed corruption cases) achieved by the classifier. The classifier is better as it approaches the NW corner of the plot, which represents the situation where the model classifies all cases correctly. Notice that the 45-degree line represents a classifier which does not outperform random guessing of the final classification. In the figure, each type of line represents a different outcome and the different panels represent each of the three models and the two strategies that we used to deal with the imbalance in the outcome variable.

Two observations are in order. First, the performance of all models is high, although the GBM has slight advantage over the other two models for all three outcome variables, especially if we consider the SMOTE approach to correct for outcome imbalance. In addition, it is clear that all models perform much better to predict an investigation by Confecá-

maras (92%/93% AUC[12]) than at predicting an investigation by the CGR (80%/81%) or an extension of the original contract (78%/80%).

Figure 1: ROC curves for all models using the three outcome variables

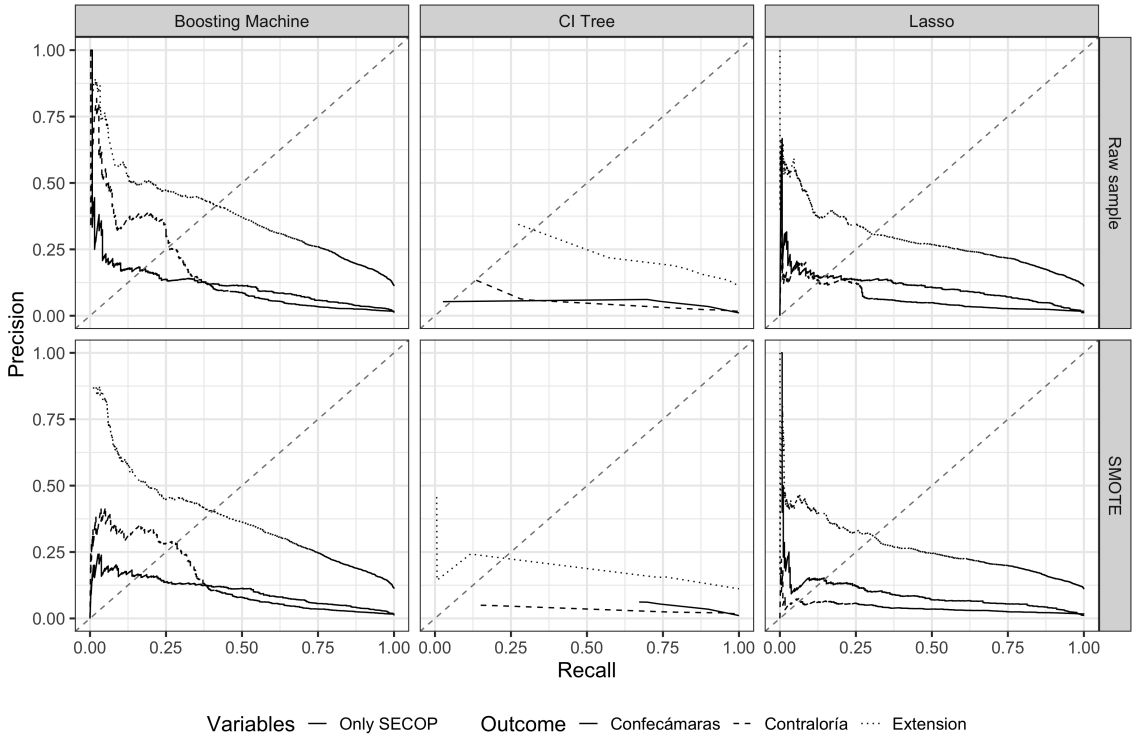


However, Figure 1 must be taken with a grain of salt. The repeatedly literature points out that ROC curves can be misleading with highly unbalanced samples (Cranmer and Desmarais, 2017). In consequence, a model that wrongly predicted all the observations as not likely to be investigated would still achieve a high accuracy. To get around this issue, the applied literature suggests that it is more informative to evaluate the models in the precision-recall space which uses the proportion of the cases that are predicted as corruption that are indeed corrupt and not the false positive rate (Cranmer and Desmarais, 2017). The precision-recall curves for the same models displayed above are shown in Figure 2. In this case, better performance is achieved by points near the NE corner. The story told by Figure 2 diverges significantly from what we saw in Figure 1. Indeed, the precision-recall curves suggest that the models predicting an investigation by Confecámaras tends to flag many more investigated cases than actually exist. The CGR model is much more equilibrated and the best performance is achieved by our model predicting extensions to

[12]The AUC, or Area Under the Curve, measures the area underneath the ROC curve and is a commonly used summary statistic of the performance of the model with a binary outcome.

contracts, for all three models and outcomes and regardless of how we deal with the unbalance in the outcome variables. Same as before, we observe that case weights and SMOTE produce moderately similar results and that the single tree model significantly underperforms the other two approaches. We also have evidence that the structure captured by the lasso and the conditional inference tree may not be sufficiently rich, compared to the GBM. In fact, the conditional inference tree seems very prone to underpredict investigations.

Figure 2: Precision-recall curves for all models using the three outcome variables

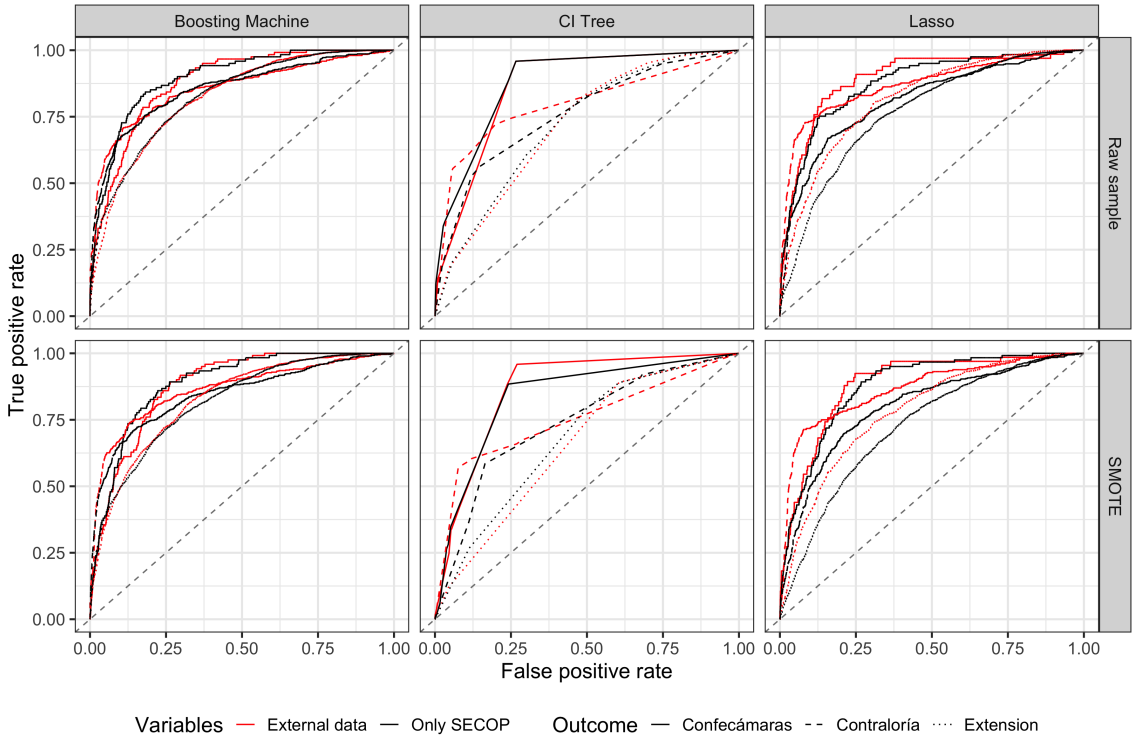


6.1.2 Municipios

We show here the results of running the same set of models as in Section 6.1.1 but for a sample of contracts at the local level (*municipio*). For these cases we can take advantage of additional information such as the economic, social, and political characteristics of the *municipio*, which in turn can tell us something about structural conditions that make investigations more likely. Same as above, we show the ROC curves (Figure 3) and the precision-recall curves (Figure 4), but now for models using two different datasets. In color red we represent the models that rely exclusively on the SECOP database, i.e., models comparable to those presented in Section 6.1.1. In black, we represent the results of models

that use a socioeconomic characteristics of the *municipio* in addition to SECOP data. Same as before, the figure represents three different modeling approaches and two different ways of managing outcome unbalance.

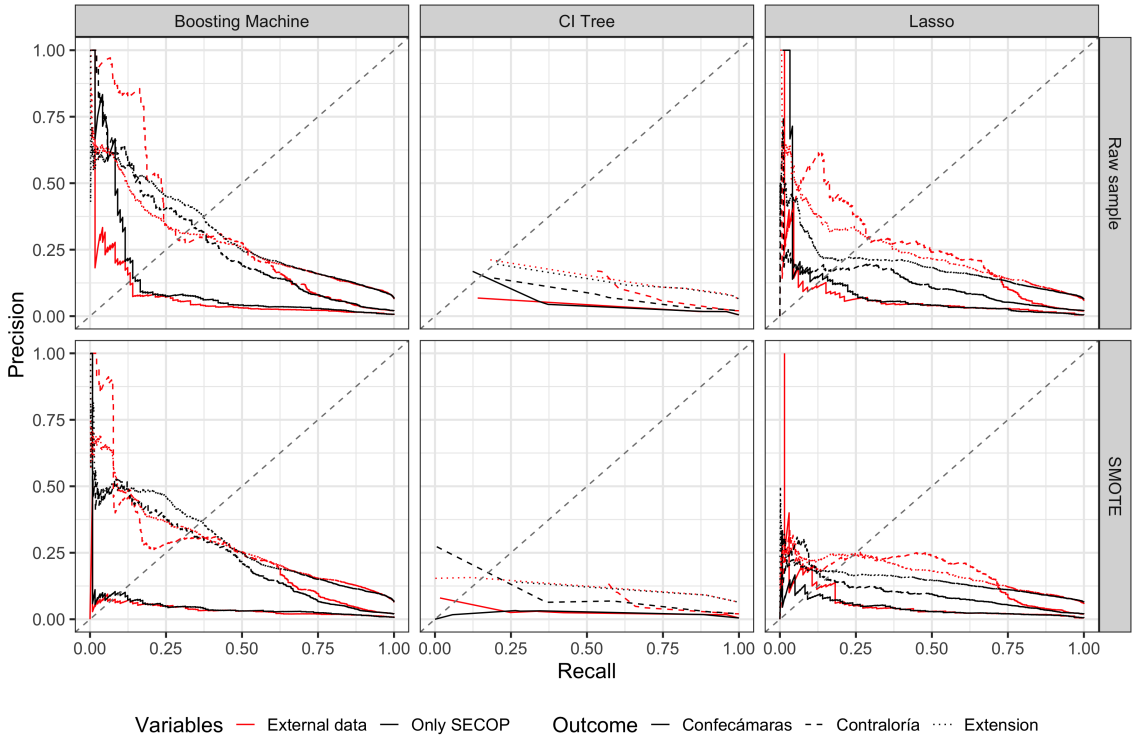
Figure 3: ROC curves for all models using the three outcome variables



The same results described above appear here: the ROC (Figure 3) tends to overestimate how well we are able to predict the cases investigated by Confecámaras relative to the CGR, and the precision-recall curves (Figure 4) sort the models in the opposite order as the ROC. Also, the GBM and the lasso show superior performance to the conditional inference tree, although not by much.

The most interesting result has to do with the value of using additional data in the predictions. We find a weak to null effect of using auxiliary data in the performance of the model with maybe the exception of the lasso which takes more advantage from having additional information about the municipio. This result is not surprising given that that the lasso imposes a very simple linear structure while the other models are more suitable to exploring transformation of the original data, like for instance, interactions.

Figure 4: Precision-recall curves for all models using the three outcome variables



6.1.3 Turning probabilities into categories

The ROC and the precision-recall curves give us a look into the global performance of the model for all possible cutoff probabilities. However, in practical applications, the decision-maker is typically interested in the predicted classification of each contract in order to decide which cases should receive further attention and resources. Different cutoff points and decision rules for how to turn a given predicted probability can be mapped to preferences over the relative weight of false positives and false negatives. For instance, agents may prefer to maximize the probability of detection ensuring that the model is able to flag all potential cases of corruption (which in turn implies a goal of minimizing false negatives). Similarly, it is also reasonable that the decision-maker wants to ensure that the model only flags cases that are very likely to need further attention, which translates into a goal of minimizing false positives.

In this setup, we argue that false positives carry a more direct and measurable financial burden. False negatives imply that the model is not indicating the agency to take any action in a case that can result corruption. False positives are associated instead with potentially unnecessary inspections that consume resources from each agency in terms of

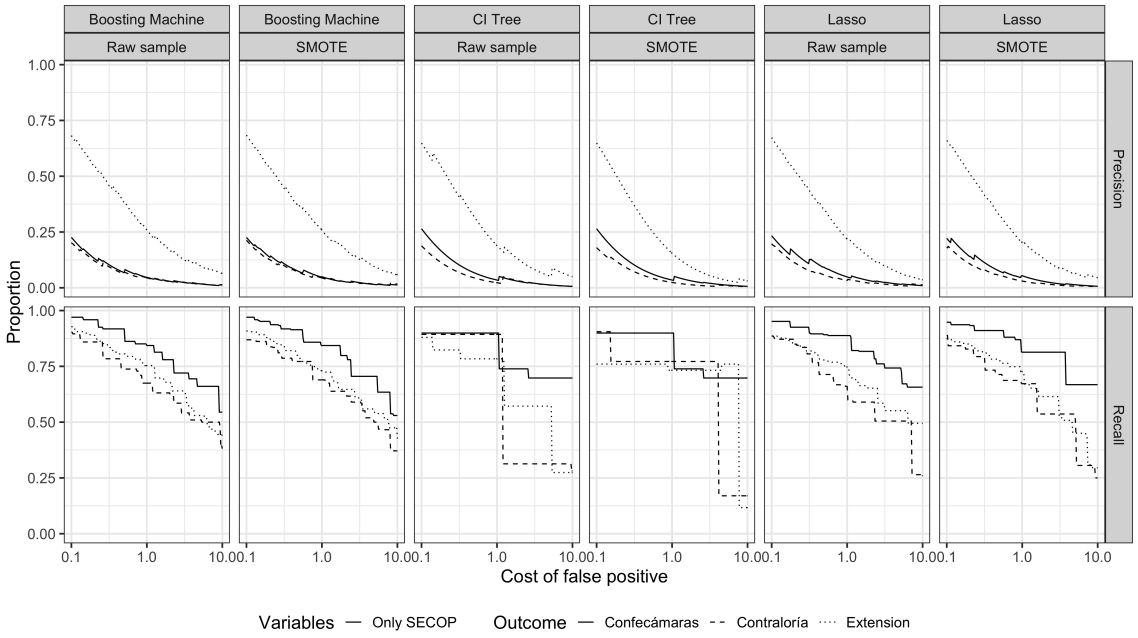
time, money, and attention. To the extent that the actions of the agency are limited by its budget, a reasonable goal is to ensure they can track as many cases as possible within their available resources. Ensuring that the pool of cases that are flagged is as small as possible and contains the minimal number of false positives seems a desirable target. Based on that consideration, in Figure 5 we simulated precision and the recall of the models for scenarios where the false positive costs the decision-maker between 0.1 and 10 times as much as a false negative.

As expected, as one increases the penalization for false positives, the precision associated with a model decreases, but it also causes to decrease the recall. The interesting observations comes from comparing the speed at which the precision and recall curves change when we place more weight on avoiding false positives, and also the different effect of this weight on the models for each of the outcomes. Indeed, the effect on the precision is substantially unaffected once false positives become more costly than false negatives, but the effect of the cost is almost linear on the recall. In other words, given the relatively low precision of the models, the actual price of avoiding false positives is mainly paid in fewer corrupt cases correctly flagged. The trade-off seems more pronounced for the case of our extension outcome, largely because the model starts with relatively high precision when false positives are cheap to flag. If we focus on the lasso and GBM models, given the poor performance of the conditional inference tree, we see that one unit increase in the cost of a false positive translates into about 1.5 and 2 points reduction in precision for an investigation by Confecámaras and the CGR and almost 6 points drop for the precision of the extension model, while the recall falls between 3.4 and 5.2 points depending on the outcome.

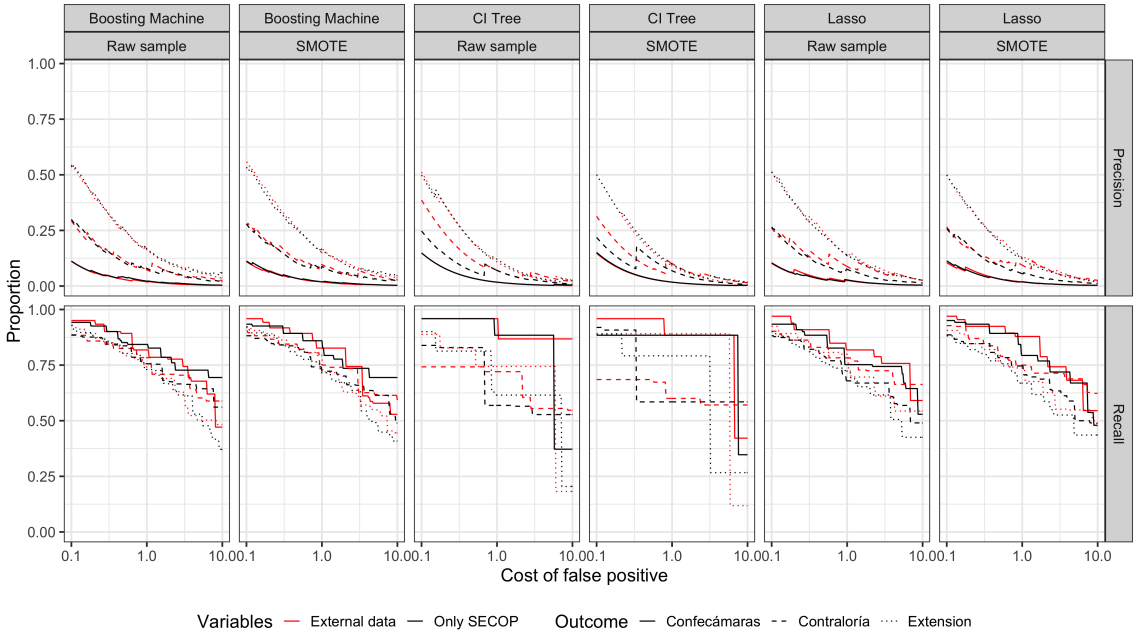
6.2. Model interpretation

The model that we found to perform the best in the previous pages, the GBM, does not lend itself to an easy interpretation of the results in the same way we would do with a parametric model like the lasso or even with a simpler model like a decision tree. We can still estimate quantities that allow us to get a better insight into what the predictive models are capturing. We are interested in two separate types of interpretations of the model. On the one hand, and from the perspective of the final user, we are interested in knowing which variables are relevant in order to predict the final outcome of a given contract. From this point of view, we are interested in discriminating among the many covariates of the model to find a subset that is not spuriously related to the outcome. Given that set, we are also interested in knowing in which way different values of the "relevant" variables affect the likelihood of the outcome. In this case, we are interested in the direction of the effects to assess even, if roughly, an approximation to the marginal effect of each value on the likelihood of malfeasance investigation.

Figure 5: Precision-recall trade-off for all models



(a) All cases

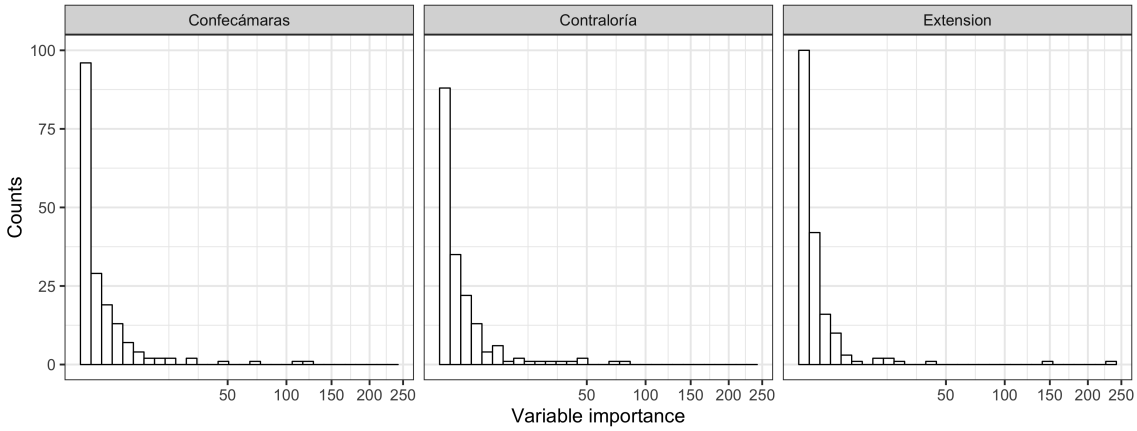


(b) Municipios only

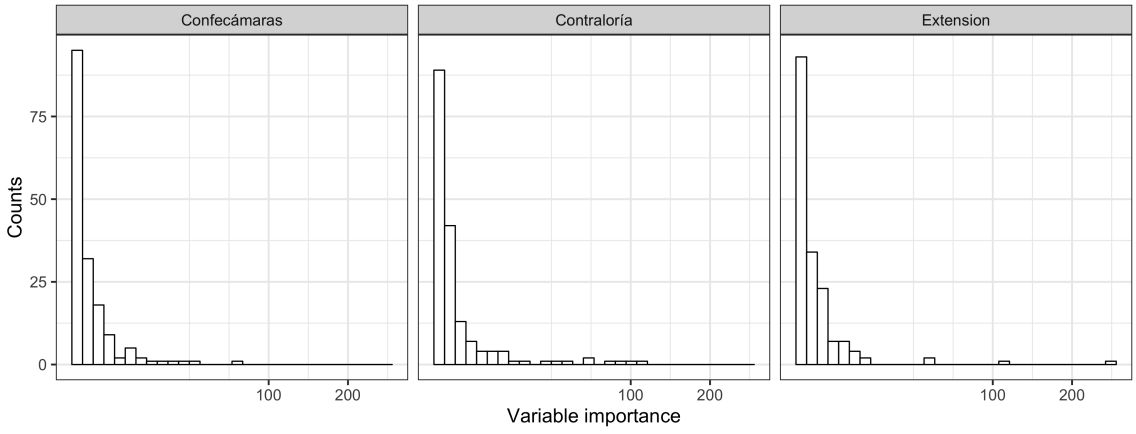
In the case of a GBM, a common approach to measure the *importance* of each variable to the model is to measure it by the reduction in the predictive capacity of the model when a given variable is permuted (Friedman et al., 2001b). In our case, and as shown in Figure 6a, a large majority of the variables are dropped and not used for prediction. In fact, only a few variables seem to be consistently selected. Also, it seems that the model using local-level information have a smoother distribution of importance, in the sense that more variables are used in the prediction. The results in the figure speaks about the relatively small information needed for the model to produce the performance described above. This regularity holds regardless of the outcome variable. Even more, in all models just a few variables stand out and have much larger importance than the rest. It thus seems that it is a small set of factors that trigger either investigations by malfeasance or extensions to contracts which in turn create a fairly well defined type of cases.

Figure 6: Distribution of variable importance

(a) All contracts



(b) Municipios only



The identity of these top variables is shown in Figure 7. We selected there the top seven predictors ranked by their unnormalized influence. As before, we focus on variables from the SECOP database only given that, as we saw in Figure 4, the additional variables at the local level do not help much in the prediction. Some clear patterns emerge from Figures 7a and 7b. If we look at the models predicting whether the contract will end up in an extension, the two principal variables are related to the size of the project. The budget of the project as well as the expected execution period, unsurprisingly, are the two single most important variables. The cases investigated by Confecámaras seem also defined mostly by their size. The budgeted cost of the contract appears as first or second variable, and the execution period is among the top seven predictors for both all the contracts in SECOP and for those at the municipal level. Those variables are also picked by the models predicting an investigation by the CGR, but only for the set of all contracts. There is therefore some clear evidence of the likelihood of detection of malfeasance and the total size of the contract. Also, while the model predicting an extension is mainly driven by the two variables mentioned above, the contracts investigated by Confecámaras and CGR have, in general, a *smoother* pattern—although the budget stands out from the other covariates when we look at local-level contracts investigated by Confecámaras.

A number of variables related to the type of contract also trigger investigations, as shown by either outcome. Remarkably, the model picks contracts related to "transportation and storage services" (*Servicios de Transporte, Almacenaje y Correo*) and also to "management, business, and administrative services (*Servicios de Gestión, Servicios Profesionales de Empresa y Servicios Administrativos*), along with specific types of procurement vehicles like a direct contract that can only be executed by specific individual providers (Type H of direct contracting).[13] Other variables related to the process are also selected by the models with high frequency. For instance, most models select as a top predictor the waiting time between the date of the award of the contract (*adjudicación*) and the *fecha de cargue*, which indicates whether there were delays in the processing of the contract.

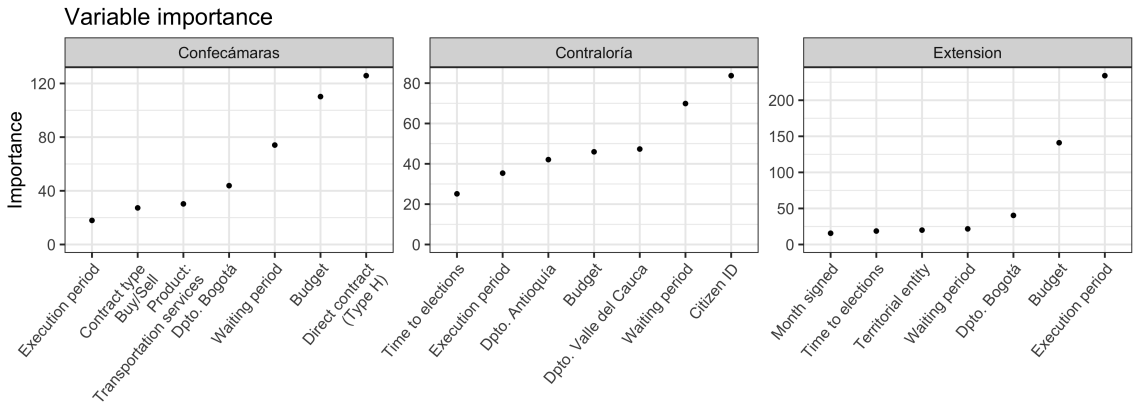
It is very significant that some of the models select among the top predictors the distance to the nearest presidential election, which can be seen as the most clear indicator of a "political-business cycle." It is especially relevant that the variable is selected by the model predicting whether the contract results in an extension, but also in the model that includes all contracts that end up investigated by the CGR.

Finally, a number of geographical predictions stand out. All models selected at least one covariate capturing specific regions (*departamentos*), like Valle del Cauca (around Cali) Cundinamarca (which surrounds the capital, Bogota), Bogota or Antioquia (which includes the second-largest city, Medellin).

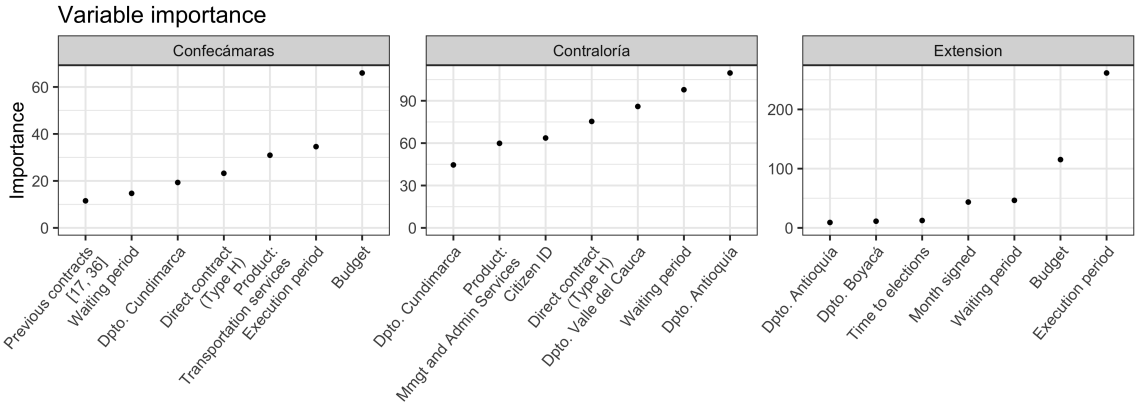
[13] Contratos de prestación de servicios profesionales y de apoyo a la gestión, o para la ejecución de trabajos artísticos que solo pueden encomendarse a determinadas personas naturales. Literal H. Numeral 4 artículo 2 Ley 1150 de 2007; artículo 81 Decreto 1510 de 2013).

Figure 7: Important variables

(a) All contracts



(b) Municipios only



Now that we have isolated which variables matter the most for the prediction, we can estimate the way in which the variables affect the outcome. We achieve that by creating "partial dependency plots," which are representations of the marginal value of a given variable after averaging over all the others (Friedman et al., 2001b). This is relevant because in Figure 7 we have isolated the variables that matter the most for the prediction, but it does not tell us the way in which they affect the probability of observing the outcome. For instance, we saw that the size of the budget and the time to execute the contract are, unsurprisingly, the most relevant variables to predict whether the contract will require an extension in either time or money. What the figure does not tell us is whether it means that extensions will be required by cheaper or more expensive projects or, potentially of

more interest, by some interaction between the two main variables. Figures 8 and 9 address that question.

They represent, for each pair of budgets and execution times, the effect on each of the three outcomes that we have discussed. Figure 8 corresponds to the model that uses a random sample of all the contracts, while 9 focuses on local-level contracts. The figures represent predictions by percentiles in both variables. They are interesting because they represent three fundamentally different types of effects of the two variables. In particular, in the case of the Confecámaras outcome, we see that while the timeline of the project may have a small effect, the effect of the budget is essentially independent: Confecámaras is more likely to collect information on more expensive projects regardless of how long they take, an effect that appears for both set of models, even if attenuated for the model using local-level data (Figure 9a).

The two variables behave very differently in the model predicting an investigation by the CGR. In particular we see that contracts in the SW and the NE are corner are much more likely to have an investigation than any others: cheap, short projects and long, expensive ones are more likely to be investigated by the agency. It is even more remarkable as it represents a clear interaction between the two variables: short projects and expensive projects are independently more likely to be investigated, but there is an independent, separate effect of having both characteristics at the same time. In the case of local contracts, the CGR seems to collect mostly information along the budget dimension with a special focus on the extremes. Although there is a trend towards more investigations in longer projects, it is noticeable that both "cheap" projects and more expensive ones are the more likely to result in an investigation.

In the model predicting whether the contract will result in an extension we observe a fairly intuitive regularity. The likelihood of an extension increases as both variables increase simultaneously. It is the combination of being both expensive and long that increases the probability of requesting additional resources, more so that being expensive or long on its own. The pattern holds for both datasets.

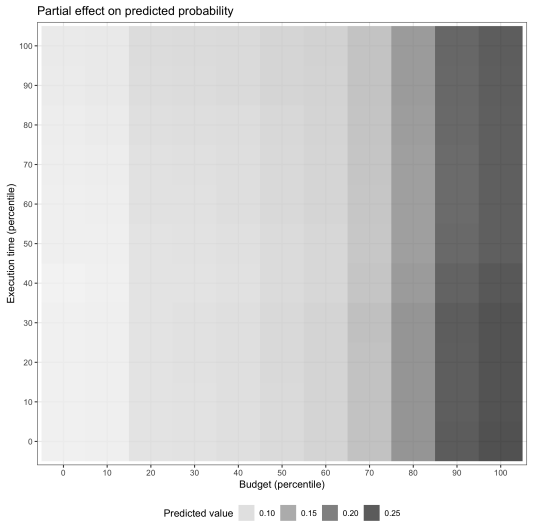
We would like to conclude this analysis by mapping the geographical distribution of risk scores in Colombia. As we mentioned above, the predicted values of our models correspond to the probability of detecting corruption for each contract. To construct departamento-level risk scores,[14] we calculate the average probability of corruption for all the contracts in a given departamento. Figure 10 shows the distribution of such risk scores for one of our models: GBM for the Confecámaras outcome, using the full sample and without the inclusion of municipality-level controls. Similar maps result for the rest of the models. Note that corruption risk is not geographically concentrated in Colombia. Resource-rich places, such as Cesar, Arauca, and La Guajira, exhibit high scores. Something similar happens

[14] A similar exercise can be conducted at the municipality level.

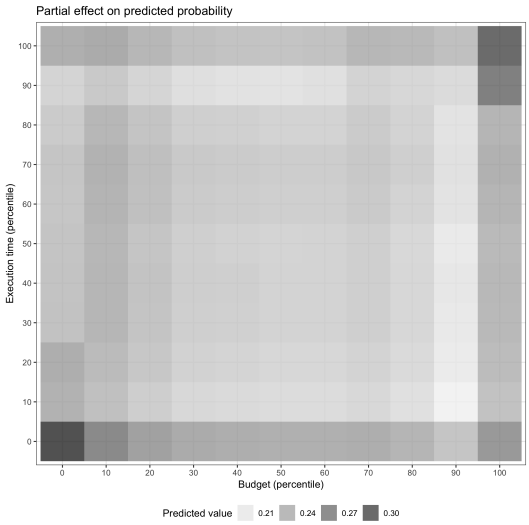
to departamentos with a high incidence of coca crops, such as Putumayo and Cauca. In general, the sidebar of the map reveals that corruption risk scores tend to be low, as the vast majority of the contracts are not classified as corrupt. However, we can still scale departamentos according to these scores. We claim that this type of analysis is useful for anti-corruption agencies that need to allocate scarce resources upon conducting audits and other monitoring activities.

Figure 8: Partial effects: All contracts

(a) Confecámaras



(b) Contraloría



(c) Extension

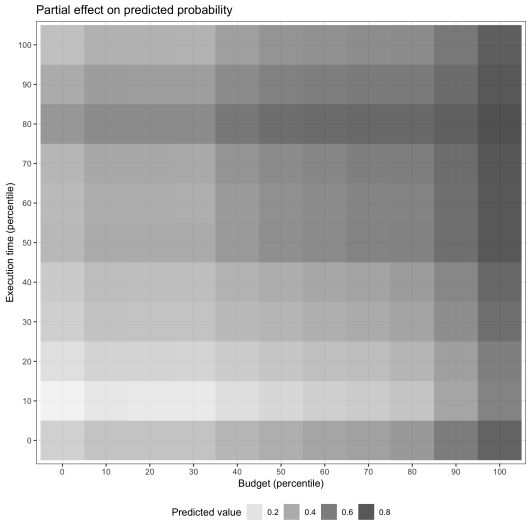
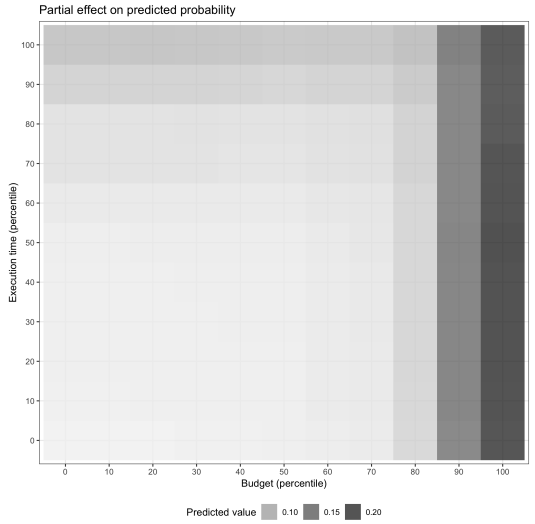
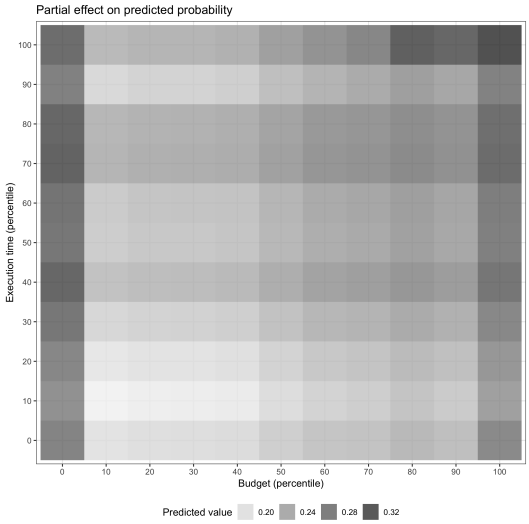


Figure 9: Partial effects: Municipios

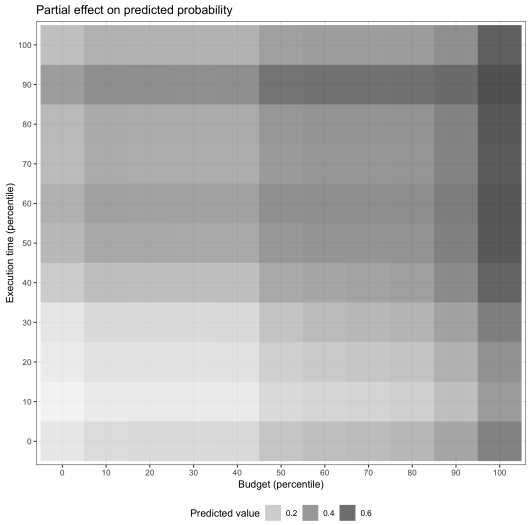
(a) Confecámaras



(b) Contraloría



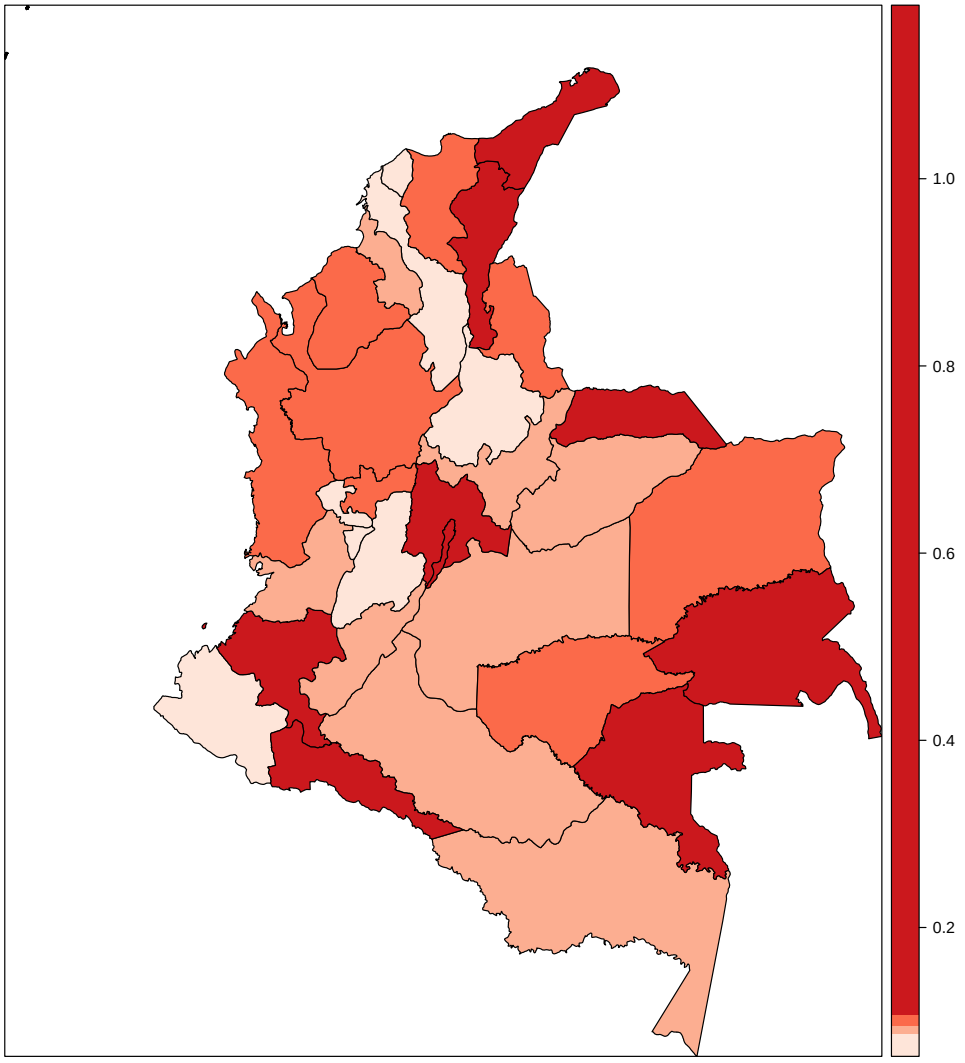
(c) Extension



7. Conclusions

Transparency is crucial to curb corruption, waste of public resources, and inefficiency. Web-based platforms to register and report public transactions have become popular in both developed and developing countries, enabling governments, through anti-corruption agencies, as well as watchdog organizations and the civil society, to use this information to monitor and prevent malfeasance. The combination of more and better information, higher storage and computational capabilities, as well as the consolidation of traditional and novel statis-

Figure 10: Geographical Distribution of Risk Scores



tical techniques to analyze this data, represent a unique setting to fight against corruption using the tools of the so-called "big data" revolution. In this paper, we have exemplified and discussed the type of analysis that practitioners and researchers can perform within this framework.

Using a unique dataset on Colombian public procurement, that documents more than 2 million contracts from 2011 to 2015, we implement machine learning algorithms—lasso, conditional inference trees and GBM—to build predictive models of faulty contracts. First, we have used different techniques to tackle the class imbalance problem of this application, given that the vast majority of contracts in our dataset are not investigated. Next, we train our models on different outcomes of corruption and inefficiency, achieving high levels of accuracy. Moreover, we are able to simulate, by varying the relative cost of having false positives, the tradeoff between precision and recall, which, in this context, implies relying on more aggressive versus more passive classifiers. Governments interested in minimizing the risk of not auditing a fraudulent contract should focus on false negatives, no matter the cost of monitoring "good" contracts. On the other hand, if auditing resources are scarce, the objective function of authorities should minimize false positives and deal with the fact that many "bad" transactions will not be discovered on time.

We argue that this type of analysis is useful from a public management perspective for at least two reasons. First, the predicted probabilities of our models represent risk scores. These may be used by the authorities when selecting which contracts they should monitor and audit. So far, in many contexts, randomness or intuition have been used to select the contracts that are audited. We hypothesize that better results may be achieved, and resources can be spent in a more efficient way if the risk scores that result from machine learning algorithms guide the authorities in these tasks. Naturally, this is an open empirical question that can be addressed using experimental methods, in collaboration with an anti-corruption agency. Contracts may be randomized, in such a way that some of them are chosen following the "traditional criteria", while others are selected based on their risk scores. Going further, it would be interesting to determine the causal effect of announcing to contractors that this procedure will be employed.

Second, the algorithms used in our models have an important characteristic. They allow us to describe which variables—and in what way—contribute more to the likelihood for a contract to be investigated or inefficient. Theoretically, this is important; and from a policy perspective, it is quite useful. It can guide the discussion upon which type of institutional reform may be more efficient in order to curb corruption. In our case, for instance, variables associated to projects, such as its size or time length, are important predictors of malfeasance. Other variables associated to the political business cycle, such as the distance between the adjudication of the contract and the nearest election, also have a high predictive value. Institutional reform focusing on procurement on election years—as it is prevalent in many contexts—seems to be crucial. In the Colombian case, there is a

grace period before an election during which certain types of contracts cannot be signed. However, a large number of pathologies cluster around the days before these deadlines.

Naturally, the discussion we have led so far does not imply that corruption will be completely thwarted if governments use public procurement data and machine learning algorithms to predict problematic contracts. This methodology still faces many challenges, as we argued in section 2. Self-selection of entities that report their information into these platforms is a major challenge. If the central government does not enforce the mandate of providing all the relevant information on each and every transaction held by a public office, it is likely that the most problematic entities will not report information truthfully. Enforcement is crucial in this case. Second, outcome variables in this realm are often difficult to obtain. Public procurement platforms may provide information on the budget and the timing of projects, but not necessarily on which of them end up being investigated. In many countries information from the judicial system is of low quality, fragmented or simply inexistent. Third, as we argued above, authorities need to prioritize and balance between precision and recall, and weigh up the costs and benefits of using aggressive or passive classifiers. Finally, it is important to remember that contractors and malfeasant public servants are not static agents that do not learn. They may perfectly anticipate which traits are more important to classify ex-ante a contract as problematic, and adjust their behavior accordingly. Hence, algorithms should also be dynamic and adapt to the new conditions and behaviors.

We would like to conclude by recalling that, so far, the quantitative analysis of public policy has focused on causal inference. This approach is understandable, as determining if a policy or program has the desired impact on its beneficiaries is crucial in many applications. But in many other cases, prediction is even more important. Resources are scarce, in such a way that forecasting certain traits or features of agents and objects is crucial to allocate these resources efficiently. This has been the case in other areas, such as security, where police forces need to anticipate with precision which zones of a city are more prone to criminal activities. In a similar fashion, public procurement is a realm in which monitoring activities and audits is costly. Hence, anticipating which contracts are riskier is essential in order to improve the quality of governance and public service delivery.

References

Ades, A. and R. D. Tella (1999). Rents, Competition and Corruption. *American Economic Review 89*(4), 982–993.

Adsera, A., C. Boix, and M. Payne (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.

Andvig, J. and O. Fjeldstad (2001). Corruption: A Review of Contemporary Research. Technical report, Bergen: Chr. Michelsen Institute.

Bardhan, P. (2006). An Economist's Approach to the Problem of Corruption. *World Development 34*(2), 341–348.

Becker, G. and G. Stigler (1974). Law Enforcement, Malfeasance, and Compensation of Enforcers. *Journal of Legal Studies 3*(1), 1–18.

Berk, R. (2012). *Criminal Justice Forecasts of Risk*. Springer.

Berton, J., P. Jaeger, and J. Grimes (2010). Using ICTs to Create a Culture of Transparency: E-Government and Social Media as Openness and Anti-Corruption Tools for Societies. *Government and Information Quarterly 27*(3), 264–271.

Besley, T. (2006). *Principled Agents? The Political Economy of Good Government*. Oxford University Press.

Bhatnagar, S. (2003). E-government and Access to Information. Global corruption report, Transparency International.

Bjorkman, M. and J. Svensson (2009). Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda. *Quarterly Journal of Economics 124*(2), 735–769.

Cawley, G. C. and N. L. Talbot (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research 11*(Jul), 2079–2107.

Chaudhury, N., J. Hammer, M. Kremer, K. Muralidharan, and H. Rogers (2006). Missing in Action: Teacher and Health Worker Absence in Developing Countries. *Journal of Economic Perspectives 20*(1), 91–116.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research 16*, 321–357.

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li (2018). *xgboost: Extreme Gradient Boosting*. R package version 0.71.2.

Chong, A., A. de la O, D. Karlan, and L. Wantchekon (2015). Does Corruption Information Inspire the Fight or Quash the Hope? A Field Experiment in Mexico on Voter Turnout, Choice, and Party Identification. *Journal of Politics 77*(1), 55–71.

Condra, L., M. Callen, R. Iyengar, J. Long, and J. Shapiro (2016). Damaging Democracy? Security Provision and Turnout in Afghan Elections. Working Paper.

Cranmer, S. J. and B. A. Desmarais (2017). What can we learn from predictive modeling? *Political Analysis 25*(2), 145–166.

Dal Bo, E. and M. Rossi (2007). Corruption and Inefficiency: Theory and Evidence from Electric Utilities. *Journal of Public Economics 91*(5-6), 939–962.

DiRienzo, C., J. Das, K. Cort, and J. Burbridge (2007). Corruption and the Role of Information. *Journal of International Business Studies 38*(2), 320–338.

Dufflo, E., R. Hanna, and S. Ryan (2012). Incentives Work: Getting Teachers to Come to School. *American Economic Review 102*(4), 1241–1278.

Ferraz, C. and F. Finan (2008). Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes. *Quarterly Journal of Economics 123*(2), 703–745.

Finan, F., B. Olken, and R. Pande (2016). *The Personnel Economics of the State*. Elsevier.

Friedman, J., T. Hastie, and R. Tibshirani (2001a). *The elements of statistical learning* (1 ed.). Springer series in statistics Springer, Berlin.

Friedman, J., T. Hastie, and R. Tibshirani (2001b). *The elements of statistical learning*, Volume 1. Springer series in statistics New York, NY, USA:.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gallego, J., J. Martinez, K. Munger, and M. Vasquez (2017). Tweeting for Peace: Experimental Evidence from the 2016 Colombian Plebiscite. Working Paper.

Grace, E., A. Rai, E. Redmiles, and R. Ghani (2016). Detecting Fraud, Corruption, and Collusion in International Development Contracts: The Design of a Proof-of-Concept Automated System. In *IEEE International Conference on Big Data*.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics 15*(3), 651–674.

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., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human Decisions and Machine Predictions. *Quarterly Journal of Economics 133*(1).

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (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., S. Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection.

Kuhn, M. (2018). *caret: Classification and Regression Training*. R package version 6.0-80.

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

Lagunes, P. (2017). Guardians of Accountability: A Field Experiment on Corruption and Inefficiency in Local Public Works. Working Paper.

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 Eexperiment of the Korean Information Infrastructure, 1995-2005. *Government and Information Quarterly 26*, 567–583.

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

Lopez-Iturriaga, F. and I. Sanz (2017). Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces. *Social Indicators Research*.

Lord, K. (2006). *The Perils and Promise of Global Transparency*. State University Press of New York.

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

Mena, J. (2011). *Machine Learning Forensics for Law Enforcement, Security, and Intelligence*. Taylor and Francis Group.

Muchlinski, D., D. Siroki, J. He, and M. Kocher (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). *TheQuest 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 Administration 32*, 415–437.

Prasad, A. and S. Shivarajan (2015). Understanding the Role of Technology in Reducing Corruption: a Transaction Cost Approach. *Journal of Public Affairs 15*(1), 22–39.

Ridgeway, G. (2017). *gbm: Generalized Boosted Regression Models*. R package version 2.1.3.

Riley, S. (1998). The Political Economy of Anti-Corruption Strategies in Africa. *European Journal of Development Research 10*(1), 129–159.

Rose-Ackerman, S. (1999). *Corruption and Government*. Cambridge University Press.

Shim, D. and T. Eom (2009). Anticorruption Effects of Information 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 (Methodological)*, 267–288.

Treisman, D. (2000). The Causes of Corruption: A Cross National Study. *Journal of Public Economics 76*(3), 399–457.

Van Hulse, J., T. M. Khoshgoftaar, and A. Napolitano (2007). Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*, pp. 935–942. ACM.

van Rijckeghem, C. and B. Weder (2001). Bureaucratic Corruption and the Rate of Temptation: Do Wages in the Civil Service Affect Corruption, and by How Much? *Journal of Development Economics 65*(2), 307–331.

Wei, S. (2000). Natural Openness and Good Government. Working Paper, NBER.

West, D. (2004). E-government and the Transformation of Service Delivery and Citizen Attitudes. *Public Administration Review 64*(1), 15–27.

World Bank (2004). Making Services Work for the Poor: World Development Report. Technical report, World Bank.

World Bank (2013). Anti-Corruption: Public Sector Management. Technical report, World Bank.