WHAT PREDICTS CORRUPTION?

EMANUELE COLONNELLI[⋆], JORGE GALLEGO[†], AND MOUNU PREM[‡]

Abstract. The ability to predict corruption is crucial to policy. Using rich micro-data from Brazil, we show that multiple machine learning models display high levels of performance in predicting municipality-level corruption in public spending. We then quantify which individual municipality features and groups of similar characteristics have the highest predictive power. We find that measures of private sector activity, financial development, and human capital are the strongest predictors of corruption, while public sector and political features play a secondary role. Our findings have implications for the design and cost-effectiveness of various anti-corruption policies.

Keywords: Corruption, Prediction, Machine learning, Private sector, Brazil

Date: December 23, 2020.
We are grateful to the Stanford Institute for Innovation in Developing Economies (SEED), the Private Enterprise Development in Low-Income Countries (PEDL) Initiative by the Centre for Economic Policy Research (CEPR), the Stanford Center for International Development (SCID), the Stanford Institute for Research in the Social Sciences (IRiSS), the Abdul Latif Jameel Poverty Action Lab (J-PAL) Governance Initiative, and Universidad del Rosario for financial support. We are grateful for the comments and suggestions at the Corruption and Big Data Conference at Universidad del Rosario.

[⋆]Booth School of Business, University of Chicago. email: emanuele.colonnelli@chicagobooth.edu.
[†]Department of Economics, Universidad del Rosario. email: jorge.gallego@urosario.edu.co.
[‡]Department of Economics, Universidad del Rosario. email: francisco.munoz@urosario.edu.co.

1. Introduction

Policy makers around the world consider the fight against corruption to be one of the most important, and yet most challenging objectives of our society. In the presence of corruption, regulations tend to be inefficient (Djankov et al., 2002), businesses are held back (Fisman and Svensson, 2007; Colonnelli and Prem, 2020), mortality rates are higher (Fisman and Wang, 2015), public and social spending is wasteful (Olken, 2007, Bandiera et al., 2009), and growth is slower (Mauro, 1995).[1]

As a result, anti-corruption policies are ubiquitous. While all policies tend to focus on some mix of monitoring and punishment of illicit acts, central to all of them is the need to effectively target the anti-corruption activity. That is, curbing corruption requires the ability to *predict* where corruption is most likely to take place. Yet, while many studies have analyzed the consequences of anti-corruption programs, little is known about what predicts corruption.[2]

In this paper, we attempt to fill this gap by focusing on the unique setting provided by Brazil's national anti-corruption audit program, which generated exogenous observable snapshots of corruption levels across thousands of municipalities over time. Based on these reports, we create two discrete measures of corruption, one for municipalities that reveal levels of corruption above the median and one for municipalities in the top quartile of the empirical distribution. The latter is constructed with the aim of capturing more severe cases of corruption at the municipality level. We complement our measure of corruption with a set of approximately 150 municipality characteristics that span different features of Brazilian municipalities. In particular, we include

[1] Some theories predict that corruption may be efficient (Leff, 1964), but these theories are mostly rejected by the empirical literature and, importantly, they refer to second-best contexts.

[2] See Olken and Pande (2012), Rose-Ackerman and Palifka (2016), and Fisman and Golden (2017) for extensive reviews of the literature.

characteristics of the private and public sectors, measures of financial development, human capital, local politics, public spending, natural resources' dependency, and other municipality characteristics.

Using our rich dataset and our measures of corruption, we first train a group of popular machine learning models as well as an ensemble model to assess whether corruption can be accurately predicted. We do so by performing a 5-fold cross-validation procedure on a training set covering 70% of our data and leaving the remaining 30% for testing the models' out-of-sample performance.

Our analysis reveals that machine learning models exhibit high levels of performance. In particular, using different measures of model performance such as AUC and accuracy, we find that tree-based models, as well as the ensemble model, outperform LASSO and Neural Networks. Our results prove to be robust to the use of a continuous measure of corruption, as well as to account for class imbalance in the case of the high corruption measure.

We then move to analyze which features have a higher predictive power on corruption, finding that private sector and human capital characteristics are the ones more likely to do so. The problem with this analysis is that there could be a group of features with high predictive power, but with no particular feature having high power by itself. To account for this, we assess the importance of a group of characteristics by computing the AUC of the model for each group. We find the strongest predictors of corruption to be those related to local private sector activity. Financial development and the quality of human capital are also relevant predictors, while variables related to the size and composition of the public sector, local politics, public spending, and natural resources' dependency have low predictive power.

A caveat to our analysis is that we abstract away from a causal interpretation of the estimates, as it is standard in prediction-focused studies. Machine learning models have recently been proven useful in other policy-related prediction issues (Kleinberg et al., 2015), such as security (Bogomolov et al., 2014), poverty (Blumenstock et al., 2015), money-laundry (Paula et al., 2016), and conflict (Blair et al., 2017; Bazzi et al., 2018; Mueller and Rauh, 2019). More related to our work, Lima and Delen (2020) use machine learning models to predict corruption perception for 132 countries, Lopez-Iturriaga and Sanz (2017) uses aggregate data and newspaper evidence from Spanish provinces to predict corruption, while Gallego et al. (2019) studies malfeasance in public procurement contracts in Colombia. Also, in the context of the recent COVID-19 pandemia, Gallego et al. (2020b) use a predicted index of corruption using machine learning to study how in places of higher predicted corruption, public sector inefficiencies and corruption can emerge in the face of a large need of expenditures as the one observed during this pandemia. Ash et al. (2020) also study the problem of predicting corruption in a context similar to ours.

Our contribution to this recent literature on the use of machine learning models in social science is twofold. On the one hand, the ability to accurately predict corruption can inform national anti-corruption policies worldwide, and help improve cost-effectiveness in a notoriously challenging and costly area to tackle. On the other hand, our results on what specific predictors matter the most shed light on the key role played by the *private* sector in the fight against corruption, which instead tends to be mostly focused on initiatives targeting the *public* sector (Hanna et al., 2011).

2. Conceptual Framework

2.1. **Corruption, Moral Hazard, and Machine Learning.** An important strand of the literature on corruption has understood this phenomenon as an agency problem (Besley and McLaren, 1993; Mookherjee and Png, 1992; Banerjee, 1997; Acemoglu and Verdier, 2000; Dabla-Norris, 2002; Aidt, 2016). Under this approach, it is often assumed that a benevolent government, enacting as the principal, needs to delegate many of its most important tasks and duties to self-interested bureaucrats. The goals of these two types of actors need not be aligned, as bureaucrats want to obtain personal gains from their activities (Aidt, 2016), while governments seek to correct market failures and maximize social welfare (Banerjee, 1997). Corruption arises when the principal cannot perfectly monitor the actions of its agents. Bureaucrats may exploit information asymmetries, accept bribes, and engage in other forms of misgovernance, anticipating not being caught by the principal. As modeled by Acemoglu and Verdier (2000), government interventions need bureaucrats in order to collect information and implement policies, but bureaucrats are self-interested and hard to monitor perfectly.

Consequently, an evident implication of this framework is that strategies aiming to reduce information asymmetries between governments and bureaucrats, or at least affecting the agents' beliefs regarding these asymmetries, may be effective in curbing corruption. In other words, the right combination of monitoring and punishment may serve as a disciplining device (Becker and Stigler, 1974). Top-down monitoring, in which higher-level officials monitor lower-level bureaucrats, represents a popular method of accountability (Olken, 2007). In this context, anti-corruption programs based on (random) audits, like the one used in Brazil, aim to tackle corruption at least through two channels: first, as mentioned above, audits increase the amount of information available to the principal related to the agents' actions. Therefore, audits enhance the government's observability. Second, given the randomness of the process, the threat of being audited should have an effect on agents' beliefs and expectations of the probability of being caught and punished. In fact, recent evidence shows that audits are useful to curb corruption (Avis et al., 2018; Colonnelli et al., 2020a) and boost economic activity (Colonnelli and Prem, 2020).

However, audits are not a flawless strategy. In particular, resources are scarce and information may be so voluminous that anti-corruption agencies may easily get overwhelmed by the amount of available data. In this context, technological innovations in general, and predictive models in particular, may represent a positive shock on monitoring capacity. Random audits may be an effective strategy to alter bureaucrats' beliefs of the probability of being scrutinized, but at the expense of allocating scarce resources in an inefficient way. Municipalities in which malfeasance is less likely may end up being audited, and vice versa. Consequently, machine learning models are useful because they allow agencies to identify, ex-ante, places in where the likelihood of corruption is higher (Gallego et al., 2019). In fact, as we show below, the models that we estimate in this paper are quite accurate in predicting corruption, suggesting that these strategies are useful if the goal is to reduce information asymmetries precisely in those places in which it would be most harmful.

2.2. **The Role of the Public and the Private Sectors.** A common feature of many anti-corruption programs that have been implemented and studied in recent years is that they target the incentives faced by bureaucrats to engage in malfeasance (Olken and Pande, 2012). Public sector wages may be a direct mechanism to discipline agents, as better-paid officials could face lower incentives to misbehave (Van Rijckeghem and Weder, 2001; Rauch and Evans, 2000; Di Tella and Schargrodsky, 2004). Higher salaries, coupled with civil service reform (Xu, 2018), may also attract better-qualified people into the public sector (Ferraz and Finan, 2011). Other pecuniary and non-pecuniary mechanisms may work as well, in such a way that additional benefits, conditioned on observable performance indicators, may discipline public officials (Glewwe et al., 2010; Muralidharan and Sundararaman, 2011; Duflo et al., 2012).

Other strategies that directly tackle the probability of being detected and punished have been implemented and studied as well. In addition to the top-down accountability strategies represented by audits that were discussed above (Olken, 2007; Ferraz and Finan, 2008; Avis et al., 2018), some other forms of monitoring have been promoted, such as grassroots participation (Bjorkman and Svensson, 2010). Moreover, both top-down and bottom-up accountability, which heavily depend on available information, may be enhanced by other mechanisms such as transparency (Djankov et al., 2010; Banerjee et al., 2012; Dunning et al., 2019), the media (Reinikka and Svensson, 2005; Ferraz and Finan, 2008), and technology (Lewis-Faupel et al., 2016; Gallego et al., 2019; Enikolopov et al., 2018).

However, these strategies reveal that the recent fight against corruption has overwhelmingly focused on bureaucrats and the public sector. Implicitly, it is assumed that features of the government, public bureaucracy, local and national level politics, electoral competition, among others, constitute the main predictors of corruption. However, corruption involves *quid pro quo* arrangements, and the role of politics in the private sector might be particularly relevant (Colonnelli et al., 2020b). In fact, early cross-country studies (Laffont and N'Guessan, 1999; Svensson, 2005) underscore the importance of economic variables directly related to entrepreneurship, such as openness and competition, upon explaining corruption. Our analysis represents an significant contribution on this front, for at least two reasons: first, our rich micro-data allows us to incorporate into the analysis of what predicts corruption an important set of features characterizing the private and financial sectors of Brazilian municipalities. Second, we use cutting-edge methods to quantify the *predictive power* of the different dimensions that may affect the levels of malfeasance encountered in the country. Surprisingly, we find that features associated with the public sector, local elections, and public spending, rank low in terms of their predictive importance, compared to variables related to the private sector and financial development.

3. Background

In May 2003, under the administration of Luis Inácio Lula da Silva, the Brazilian central government launched a large anti-corruption program to fight the rampant corruption in the waste of public resources by local governments. The program consisted of 39 rounds of randomized audits of municipalities' expenditures—with replacement—over the 2003-2014 period, followed by anti-corruption enforcement activities, such as the suspension of corrupt public officials and politicians.

The audits are conducted by the Office of the Comptroller General (Controladoria Geral da Uniao (CGU)), which is the federal agency responsible for ensuring the transparent use of public funds and is considered the main anti-corruption body in Brazil. At each audit round, approximately 60 municipalities were randomly selected, with replacement.[3] As of 2014, more than 99% of Brazil's 5,570 municipalities were eligible, and 1,881 had been selected at least once. Only municipalities below a certain population threshold were eligible for the program, and state capitals were excluded.

The audit process begins immediately after the random draw, with the federal CGU office detailing the audit to the various CGU state offices by means of a number of inspection orders. The audits investigate how the federal transfers from the central government to the municipality are spent and focus mostly on the previous three years. During an intense few weeks of field works, the auditors analyze all relevant documents and receipts related to the spending of federal funds, interview local people, bureaucrats, and other relevant parties, solicit direct anonymous complaints about malfeasance, and take pictures to document the quality of public service delivery. Following this fieldwork, the auditors write a detailed audit report following the meticulous instructions from the federal CGU. These publicly available reports can span up to 300 pages and include organized analyses of all the information gathered during the weeks-long audit.

[3] The randomization is linked to the draw of a popular national lottery. The implied audit probability in any given round, which is constant within a state, is therefore quite low (1% within a round, and 3% within a year). Additionally, there is a small exception to the random draw with replacement, as municipalities cannot be selected if they were selected in one of the previous three rounds.

4. Data

4.1. **Measuring Corruption.** Measuring corruption is challenging, and typical sources of information such as self-reported perceptions or malfeasance cases covered by the media tend to suffer from severe measurement error (Sequeira, 2012). To alleviate these concerns, we focus on Brazil's anti-corruption program discussed in the previous section. Since municipalities are not able to anticipate the audit, and because of the uniform criteria adopted by highly paid federal auditors in the auditing process, this setting is uniquely well-suited to the measurement of our main outcome variables.

Our primary measures of corruption intensity in a municipality are observed the year the audit takes place using administrative data collected by the anti-corruption federal agency that oversees the program, namely CGU. Out of 5,570 municipalities in Brazil, 1,084 have been randomly selected for at least one audit during the 2007-2014 audit period we study. We focus on two binary definitions of corrupt municipalities, constructed using the share of the total number of irregularities over the size of the municipality.[4] Irregularity cases are extremely heterogeneous, ranging from cases of mismanagement in the allocation of public funds to outright bribery in government procurement. We consider corruption to be any case of moderate or severe irregularity as defined by CGU. "Corrupt" ("Highly Corrupt") municipalities are those with an above-median (top quartile) share in the distribution of corruption across all municipalities audited.

4.2. **Covariates.** We augment our analysis with granular data on local characteristics at the municipality-year level that comes from multiple confidential and publicly available sources.

We use 147 covariates that we group into eight categories: i) private sector includes different measures of economic activity and sectoral distributions, ii) public sector

[4] Municipality size is computed as the number of business establishments in the municipality.

features include the size, relative importance, and wages of public officials, iii) financial development includes measures of credit-related variables from public and private banks, iv) human capital includes measures of education and access to it, v) public spending includes different types of spending as well as local procurement variables, vi) local politics includes variables of political competition and alignment with the central government, vii) natural resources' dependency includes the relevance of different natural resources, and finally viii) local demographics include variables related to income distribution, health statistics, and crime.

The data sources and exact definitions of each variable are reported in Table 1. All variables, except the few in the Decennial Census, are measured as averages in the three years prior to the audit.

5. Machine Learning Models

In this section, we describe the machine learning models used to predict corruption as well as the training procedure and the different measures we use to assess the performance of the different models.

5.1. **Models.**

In order to predict municipality-level corruption, we train a set of popular machine learning models, which include "Random Forests," "Gradient Boosting," "Neural Networks," and "LASSO." Each model has weaknesses and strengths, and therefore we also rely on an ensemble model that combines the predictive capabilities of all individual models to optimize performance (Friedman et al., 2001). We ultimately let the data inform which model is best suited for this application based on out-of-sample performance.

5.1.1. *Lasso.*

The LASSO regression, first introduced by (Tibshirani, 1996), is similar to a logistic regression, but adds a penalization term based on the sum of the absolute values of the coefficients. This penalization term aims at shrinking the parameters towards zero. Hence this estimator is similar to a logit model, but it is more parsimonious, adding only those variables that are relevant predictors. One of the advantages of this model is that it is simple and less prone to over-fitting. However, it is incapable of identifying complex relationships between the predictors and our outcome variable, i.e., corruption. The tuning parameter in the cross-validation is the weight of the penalization term in the objective function (λ), which is optimized over a grid of potential values.

5.1.2. *Random Forests.* Random Forests are ensembles of many decision trees, where each one of them is a sequence of rules that divides the sample into sub-groups (called leaves) based on certain variable cutoffs. The prediction for each leaf, in the case of a classification task, is the most common outcome for the trained observations on that leaf, and the trees are fit so as to maximize the information gain of the resulting partitions of the data. Each tree in a Random Forest is constructed by sampling a random subset of the training data and a random subset of the predictors. Each of these trees generates a prediction, and the overall prediction of the Random Forest is the average (or the majority) of the predictions among all trees. In this application, we keep fixed the number of fitted trees (500) and use cross-validation to determine the optimal number of features available in every node.

5.1.3. *Gradient Boosting Machine.* Gradient Boosting Machines (GBM) are ensembles of weak learners, in this case, decision trees. Under boosting, classification algorithms are sequentially applied to a reweighted version of the training data (Friedman et al., 2000). GBM is a variant of Random Forests, in which trees are not fitted randomly nor independently. Instead, each tree is fitted sequentially to the full dataset, in such a way that the weaknesses of trees are identified by using gradients in the loss function, allowing subsequent predictors to learn from the mistakes of the previous ones. In other words, a gradient descent procedure is used to minimize the loss when adding new trees. Consequently, as opposed to Random Forests, observations are not selected via bootstraping, but as a function of past errors. In this way, the addition of each tree offers a slight improvement in the model (Freund et al., 1999). In our models, we keep fixed the learning rate (shrinkage parameter) and the minimum number of observations in the terminal nodes to avoid overfitting, and use cross-validation to determine the optimal number of trees and the interaction depth.

5.1.4. *Neural Networks.* Neural networks model the relationship between input and output signals through models that mimic the way biological brains work. In particular, neural networks are composed of three basic elements: an activation function, that for each neuron, transforms the weighted average of input signals (predictors) into an output signal; a network topology, which is composed by the number of neurons, layers, and connections used by the model; and a training algorithm, which determines the way in which connection weights are set with the task of activating or not neurons as a function of the input signals. This process determines the final prediction of the model. The most common activation functions include the logistic sigmoid, linear, saturated linear, hyperbolic tangent, and Gaussian (Radial Basis) functions. In the end, the process entails an optimization problem in which the optimal weights of the input signals are determined for each node. In this analysis, we keep fixed a logistic activation function and use cross-validation to determine the optimal number of units in the hidden layer (size) and the regularization parameter (decay).

5.1.5. *Super Learner Ensemble.* Ensembles are collections of predictors which are grouped to each other in order to give a final prediction. It is usually the case that ensembles—as they result from the combination of different models—perform better than their individual components. For our analysis, we use the Super Learner ensemble method developed by Polley et al. (2011), which finds an optimal combination of individual prediction models by minimizing the cross-validated out-of-bag risk of these predictions. It has been shown that the Super Learner performs asymptotically as well as the best possible weighted combination of its constituent algorithms (Van der Laan et al., 2007). We use the Super Learner models not only to stack the individual predictions, but also to test for the relative importance of different groups of variables to predict corruption.

5.2. **Training and Testing.** We use an indicator variable for corruption in year t as our variable of interest, while all predictors are measured as averages between the year $t-1$ and $t-3$, and in the case of census variables, they are all measured in 2000. In this way, we end up with a cross-sectional dataset with all the municipalities that were audited at least once between 2007 and 2014. For those audited more than once, we only use the first audit. In order to train our models, we conduct the following procedure:

(1) We divide our dataset into 70% as our training set and 30% as our testing set.

(2) In our training set, we perform a 5-fold cross-validation procedure in order to train our models and choose the optimal combination of parameters. This method divides the training set into five different equal size samples at random. Then, a model is fit in four subsamples and then test it in the remaining one. We repeat this procedure for each of the five subsamples, so each one of them ends up being a validation set, and for each of the values of the tuning parameter grid of each model. Then, the best performing parameters are chosen.

(3) The previous step is repeated 10 times with different random partitions. Hence, we obtain 10 "optimal parameters" and we use as our optimal parameter the average of them. For the case of integer parameters, we round it to the closest integer.

(4) Using these optimal parameters, we assess the performance of our models in the testing set that has never been used for training purposes.

We standardize the data by the mean and standard deviation of the training set. Table 2 shows the optimal parameters of our training procedure for each of our models.

5.3. **Assessing Models' Performance.** Once we have calibrated our model following the cross-validation procedure explained above, we compare the performance of the different models using the test set. We use as a first performance measure of interest the area under the ROC (Receiver Operating Characteristic) curve (AUC). This is a measure of the trade-off between the true positive rate and false positive rate, as we vary the discrimination threshold. It can also be interpreted as the probability that, if we randomly select two observations, they will be correctly ordered in their predicted risk of corruption, i.e., the probability that the municipality at a greater risk for corruption is assigned a higher probability of corruption. We also present each model's level of *accuracy*, which corresponds to the proportion of municipalities correctly predicted as corrupt; models' *precision*, which is the proportion of positive identifications that are correct (or true positives over true positives plus false positives); models' *recall*, which is the proportion of actual positives identified correctly (true positives over true positives plus false negatives), and models' *F1*, which is the harmonic mean of precision and recall.

5.4. **Identifying Best Predictors.** To identify the municipality characteristics that best predict corruption, we first use *covariate* importance measures. For tree-based models, importance is measured as the information gain, or the homogeneity in the resulting partitions of our set of municipalities, achieved when splitting on each variable. In the procedure that we implement, importance is measured on a scale from 0 to 100, in such a way that each variable's information gain is expressed relative to the variable with the highest information gain. Hence, the most important predictor receives a score of 100 according to this scale and the scores start to decrease for the remaining variables. For the LASSO model the importance is determined by the estimated coefficients of the regression, where larger parameters (in absolute value) correspond to higher importance. In the case of neural networks, importance is determined by the weights that connect neurons within the network.

We then move to the analysis of the predictive performance of subgroups of related predictors in order to understand which categories matter the most. It may be the case that some groups do not have one particular variable that highly predicts corruption, but that the group as a whole has high predictive power. We perform this analysis in the following way. We estimate models including each category individually (i.e., excluding all variables that are not part of it) and compute the resulting AUC for the group. Then, we rank them according to their AUC, and compare the computed AUC with a 50% level, which corresponds to the AUC of a random prediction "model." The category that increases the AUC by itself the most is the model with the highest predictive power level. We compute confidence intervals at a 95% confidence level by performing bootstrapping over the test set and computing the AUC for each sample. In this way, we are able to determine if there are any statistically significant differences in AUCs across categories.

6. Findings

In this section, we present the results of our analysis. First, we focus on the overall performance of the predictive models and their robustness to alternative measures and specifications. Then, we identify the best individual and group predictors and their link to the corruption literature.

6.1. Models' Performance and the Predictability of Corruption.

Figure 1 depicts the performance of our models. Using the two primary corruption measures of "Corrupt" (Panel A) and "Highly Corrupt" (Panel B) municipalities, we present the

ROC curves of each individual model and the ensemble model: the models perform extremely well in predicting both corruption measures. Table 3 reports the AUC levels for every model, which ranges from a minimum of 0.95 (0.94) for Neural Networks to a maximum of 0.98 (0.99) for Gradient Boosting and the ensemble model when predicting "Corrupt" ("Highly Corrupt") municipalities. Generally, AUC levels of 0.8 and above are considered excellent.

Overall, in terms of individual models, Figure 1 shows that our tree-based algorithms, namely Gradient Boosting and Random Forest, outperform LASSO and Neural Networks. We find this to be the case not only in terms of AUC levels, but also concerning precision, recall, and F1, as it is evident from Table 3. Not surprisingly, the ensemble model performs best, as it is constructed by optimizing the weights of each individual model.

In sum, these results suggest that by using fine-grained information from Brazilian municipalities, we are able to predict which areas exhibit higher levels of corruption. This is an important result from a policy perspective, as recent evidence shows that anti-corruption audits are effective tools to curb corruption (Avis et al., 2018) and boost economic activity (Colonnelli and Prem, 2020). However, at the same time they are expensive to conduct and are therefore restricted to a limited number of target areas. Risk scores estimated through machine learning models may help anti-corruption agencies optimize their resources, in such a way that audits may target those places in which information asymmetries are predicted to be more harmful.[5] In fact, recent efforts in European countries are being conducted in this direction (Petheram et al., 2019).

[5] Indeed, motivated by value-for-money concerns, Brazil's anti-corruption agency recently moved to a semi-randomized audit program, where previous audit results are used with the goal of predicting the highest-risk municipalities to target. This strategy is common across several Supreme Audit Institutions around the world.

6.2. **Robustness and Additional Analyses.** We now present alternative specifications to test for the robustness of our main results. Specifically, we present the model performance for a continuous measure of corruption, i.e., the number of cases over the number of establishments. We estimate the continuous versions of our four models and compare their performance with a (naive) baseline model, in which the prediction is simply the mean value of our outcome variables. To measure performance, we use traditional metrics such as the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), and the in-sample R-squared (see Table 4). Overall, our machine learning models perform better than the baseline case, with Random Forests and GBM usually achieving the highest levels of performance, as in the case of our discrete measure of corruption.

Additionally, we show that our findings for the "High Corrupt" dummy are robust to account for the class imbalance in the outcome. Class imbalance may be an issue when the relevant category of the outcome that we want to predict, high levels of corruption in our case, is considerably less frequent than the other category. Different methods have been proposed in the literature to deal with this problem. Given the nature of our data, we use over- and under-sampling techniques to randomly increase (decrease) the number of highly corrupt (non-highly-corrupt) municipalities. Table 5 shows that our results remain largely unchanged, suggesting that the high levels of predictive performance achieved by our original models are not driven by class imbalance.

Finally, we also estimate models for the discrete outcomes in which quadratic and interaction terms of all of our predictors are incorporated in order to account for non-linearities and more complex associations between corruption and municipality-level characteristics. In terms of model performance, the results of these estimations, available upon request, are quite similar to what we encountered for our baseline models.

6.3. **What Are the Best Predictors of Corruption?** We now move to the analysis of the individual covariates that best predict corruption. Figure 2 presents the covariate-specific importance in predicting both outcome variables of “Corrupt” and “Highly Corrupt” municipalities, and restricting the focus to the top ten features in each case. The results highlight the striking importance of a primary private sector covariate, namely the count of business establishments in the formal sector, in predicting corruption. Other important predictors are measures of market competition and human capital. These results go in line with early cross-country evidence (Svensson, 2006), which suggests that corrupt places tend to be less open to competition and regulate more the entry of firms to markets.

We also implement a variable selection procedure following Belloni et al. (2014). Table 6 presents the OLS from the doubly-robust LASSO suggested by the authors. We find that five to six variables are selected as “important” predictors, which suggests that our models are sparse. In this context, sparsity is a desirable trait, as it shows that our machine learning models are capable of simplifying a complex high-dimensional case into a simpler low-dimensional model that is easier to interpret (Hastie et al., 2015), something that conventional methods—such as OLS—will hardly achieve. This procedure allows us to determine which individual covariates matter the most and what is the direction of their correlations with corruption. In particular, these results show that private sector concentration (HHI) and the share of the construction sector are positively correlated with corruption. Other variables related to the private sector and financial development also exhibit high levels of predictive power.

Motivated by these individual ranking analysis, in Figure 3, we perform an estimation where we categorize all 147 covariates into eight groups, as shown in Table 1. Sequentially and separately, adding each group to the estimation of the ensemble model, we assess the performance of each of them as measured by the AUC. We also present confidence intervals at a 95% confidence level by performing bootstrapping over the test set.

Consistent with our analysis of individual features, we find that the *private sector* category is the strongest predictor of corruption, followed by the categories of *financial development*, *local demographics*, and *human capital* (see Panel A Figure 3).[6] The categories of *public sector*, *natural resources' exposure*, and *public spending* are less important predictors, and *local politics* is the least important one for both measures of corruption.

These results are somewhat surprising, given the overwhelming focus of both the academic and policy literature on the latter category types. For example, several studies of patronage suggest that the size of the public sector is strongly linked to corruption (Robinson and Verdier, 2013; Gallego et al., 2020a; Colonnelli et al., 2020c). Similarly, an important strand of literature has focused on the key role played by public sector compensation in curbing corruption (Di Tella and Schargrodsky, 2003; DalBo et al., 2013). Other studies suggest that elections may discipline politicians, as informed voters may punish candidates who engage in corrupt activities (Ferraz and Finan, 2008; Chong et al., 2015; Dunning et al., 2019). Moreover, the resource-curse literature (Sachs and Warner, 1995) suggests that corruption may be one of the reasons explaining why resource-rich places often exhibit lower levels of development (Shaxson, 2007). The emphasis on the role of the private and financial sectors, on the other hand, remains significantly lower (Rose-Ackerman and Palifka, 2016).

7. Conclusions

The ability to predict corruption is crucial to policy. In the context of Brazilian municipalities, we show that machine learning models and rich micro-data provide a powerful combination to accurately predict where corruption in local public spending

[6] Local demographics include a host of health and population measures, as well as other measures such as media access which do not perfectly fit into the other seven categories.

is most likely to take place. Interestingly, we find that private sector, financial development, and human capital features are the most important predictors of corruption, while public sector and political features play a secondary role.

Our findings have important policy implications that may affect how we think the fight against corruption should be held. First, the fact that our algorithms achieve high levels of performance implies that these type of methods, coupled with recent advances on the fronts of technology and transparency, represent a positive shock on governments' monitoring capacity. Audits conducted by anticorruption agencies throughout the world tend to follow heuristic rules or random assignment mechanisms. Randomness may have important effects on agents' beliefs and expectations but at the expense of generating inefficient allocations of scarce auditing resources. Our results suggest that a targeted distribution of monitoring, based on the risk scores that result from machine learning algorithms, may help governments distribute audits to places in which information asymmetries may be more harmful.

In addition, it is important to recognize that recent efforts to control corruption, such as the consolidation of top-down and bottom-up accountability mechanisms, salary and incentive-based interventions, civil service reforms, among others, overwhelming rest on the assumption that features associated with the public sector and local politics are the most important predictors of malfeasance. However, our covariate-importance metrics, both at the individual and especially at the group level, reveal that other dimensions are more important in predicting corruption. In particular, our models suggest that features associated with the private sector and financial development across Brazilian municipalities achieve the highest levels of predictive power. Hence, even though our analysis is not causal in nature, it tentatively suggests that a new generation of interventions to be tested and implemented in the future should focus more on the role that the private sector, financial institutions, competition, and markets play in enhancing or curbing corruption.

References

Acemoglu, D. and T. Verdier (2000): "The choice between market failure and corruption," *American Economic Review*, 90, 194–211.

Aidt, T. (2016): "Rent seeking and the economics of corruption," *Constitutional Political Economy*, 27, 142–157.

Ash, E., S. Galletta, and T. Giommoni (2020): "A Machine Learning Approach to Analyzing Corruption in Local Public Finances," *Available at SSRN*.

Avis, E., C. Ferraz, and F. Finan (2018): "Do Government Audits Reduce Corruption? Estimating the Impacts of Exposing Corrupt Politicians," *Journal of Political Economy*, 126, 1912–1964.

Bandiera, O., A. Prat, and T. Valletti (2009): "Active and Passive Waste in Government Spending: Evidence from a Policy Experiment," *American Economic Review*, 99, 1278–1308.

Banerjee, A. (1997): "A Theory of Misgovernance," *The Quarterly Journal of Economics*, 112, 1289–1332.

Banerjee, A., S. Mullainathan, and R. Hanna (2012): "Corruption," Tech. rep., National Bureau of Economic Research.

Bazzi, S., R. A. Blair, C. Blattman, O. Dube, M. Gudgeon, and R. Peck (2018): "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia," .

Becker, G. and G. Stigler (1974): "Law enforcement, malfeasance and the compensation of enforcers," *Journal of Legal Studies*, 3, 1–19.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "High-Dimensional Methods and Inference on Structural and Treatment effects," *Journal of Economic Perspectives*, 28, 29–50.

Bernstein, S., E. Colonnelli, D. Malacrino, and T. McQuade (2018): "Who Creates New Firms When Local Opportunities Arise?" Tech. rep., National Bureau of Economic Research.

Besley, T. and J. McLaren (1993): "Taxes and Bribery: The Role of Wage Incentive," *Economic Journal*, 103, 119–141.

Bjorkman, M. and J. Svensson (2010): "When is community-based monitoring effective? Evidence from a randomized experiment in primary health in Uganda," *Journal of the European Economic Associarion*, 8, 571–581.

Blair, R. A., C. Blattman, and A. Hartman (2017): "Predicting local violence: Evidence from a panel survey in Liberia," *Journal of Peace Research*, 54, 298–312.

Blumenstock, J., G. Cadamuro, and R. On (2015): "Predicting Poverty and Wealth from Mobile Phone Metadata," *Science*, 350, 1073–1076.

Bogomolov, A., B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland (2014): "Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data," in *Proceedings of the 16th international conference on multimodal interaction*, ACM, 427–434.

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, 55–71.

Colonnelli, E., S. Lagaras, J. Ponticelli, M. Prem, and M. Tsoutsoura (2020a): "Revealing Corruption: Firm and Worker Level Evidence From Brazil," *Working paper.*

Colonnelli, E., V. P. NETO, and E. Teso (2020b): "Politics At Work," *Available at SSRN 3715617.*

Colonnelli, E. and M. Prem (2020): "Corruption and Firms," Mimeo.

Colonnelli, E., M. Prem, and E. Teso (2020c): "Patronage and Selection in Public Sector Organizations," *American Economic Review*, 110, 3071–99.

Dabla-Norris, E. (2002): "A game theoretical analysis of corruption in bureaucracies," in *Governance, corruption, economic performance*, ed. by G. T. A. . S. Gupta, The International Monetary Fund, chap. 5, 111–134.

DalBo, E., F. Finan, and M. Rossi (2013): "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service," *Quarterly Journal of Economics*, 128, 1169–1218.

Di Tella, R. and E. Schargrodsky (2003): "The Role of Wages and Auditing during a Crackdown on Corruption in the City of Buenos Aires," *Journal of Law and Economics*, 46, 269–292.

——— (2004): "Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack," *American Economic Review*, 94, 115–133.

Djankov, S., R. L. Porta, F. L. de Silanes, and A. Shleifer (2002): "The Regulation of Entry," *Quarterly Journal of Economics*, 117, 1–37.

——— (2010): "Disclosure by politicians," *American Economic Journal: Applied Economics*, 2, 179–209.

Duflo, E., R. Hanna, and S. Ryan (2012): "Incentives work: getting teachers to come to school," *American Economic Review*, 102, 1241–1278.

Dunning, T., G. Grossman, M. Humphreys, S. Hyde, and C. McIntosh (2019): *Metaketa I: The Limits of Electoral Accountability*, Cambridge University Press.

Enikolopov, R., M. Petrova, and K. Sonin (2018): “Social Media and Corruption,” *American Economic Journal: Applied Economics*, 10, 150–174.

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

——— (2011): “Motivating Politicians: The Impacts of Monetary Incentives on Quality and Performance,” Working Paper.

Fisman, R. and M. Golden (2017): *Corruption. What Everyone Needs to Know*, Oxford University Press.

Fisman, R. and J. Svensson (2007): “Are corruption and taxation really harmful to growth? Firm level evidence,” *Journal of Development Economics*, 83, 63–75.

Fisman, R. and Y. Wang (2015): “The Mortality Cost of Political Connections,” *Review of Economic Studies*, 82, 1346–1382.

Freund, Y., R. Schapire, and N. Abe (1999): “A Short Introduction to Boosting,” *Journal-Japanese Society For Artificial Intelligence*, 14, 1612.

Friedman, J., T. Hastie, and R. Tibshirani (2001): *The Elements of Statistical Learning*, vol. 1, Springer series in statistics New York, NY, USA:.

Friedman, J., T. Hastie, R. Tibshirani, et al. (2000): “Additive Logistic Regression: A Statistical View of Boosting (with discussion and a rejoinder by the authors),” *The Annals of Statistics*, 28, 337–407.

Gallego, J., C. Li, and L. Wantchekon (2020a): “A Theory of Broker-Mediated Clientelism,” Mimeo.

Gallego, J., G. Rivero, and J. Martinez (2019): “Preventing Rather than Punishing: An Early Warning Model of Malfeasance in Public Procurement,” Mimeo.

Gallego, J. A., M. Prem, and J. F. Vargas (2020b): “Corruption in the Times of Pandemia,” *Available at SSRN 3600572*.

Glewwe, P., N. Ilias, and M. Kremer (2010): “Teacher Incentives,” *American Economic Journal: Applied Economics*, 2, 205–227.

Hanna, R., S. Bishop, S. Nadel, G. Scheffler, and K. Durlacher (2011): “The Effectiveness of Anti-Corruption Policy,” *EPPI Centre Report*.

Hastie, T., R. Tibshirani, and M. Wainwright (2015): *Statistical Learning with Sparsity. The Lasso and Generalizations*, Taylor and Francis Group.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015): "Prediction Policy Problems," *American Economic Review: Papers and Proceedings*, 105, 491–495.

Laffont, J.-J. and T. N'Guessan (1999): "Competition and corruption in an agency relationship," *Journal of Development Economics*, 60, 271–295.

Leff, N. H. (1964): "Economic Development Through Bureaucratic Corruption," *American Behavioral Scientist*, 8, 8–14.

Lewis-Faupel, S., Y. Neggers, B. Olken, and R. Pande (2016): "Can electronic procurement improve infrastructure provision? Evidence from a large rural road program in India," *American Economic Journal: Economic Policy*, 8, 258–283.

Lima, M. S. M. and D. Delen (2020): "Predicting and explaining corruption across countries: A machine learning approach," *Government Information Quarterly*, 37, 101407.

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

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

Mookherjee, D. and I. Png (1992): "Monitoring vis-a-vis Investigation in Enforcement of Law," *American Economic Review*, 82, 556–564.

Mueller, H. F. and C. Rauh (2019): "The hard problem of prediction for conflict prevention," .

Muralidharan, K. and V. Sundararaman (2011): "Teacher performance pay: experimental evidence from India," *Journal of Political Economy*, 119, 39–77.

Olken, B. (2007): "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, 115, 200–249.

Olken, B. and R. Pande (2012): "Corruption in Developing Countries," *Annual Review of Economics*, 4, 479–509.

Paula, E. L., M. Ladeira, R. N. Carvalho, and T. Marzagao (2016): "Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 954–960.

Petheram, A., W. Pasquarelli, and R. Stirling (2019): "The Next Generation of Anti-Corruption Tools: Big Data, Open Data and Artificial Intelligence," Tech. rep., Oxford Insights.

Polley, E. C., S. Rose, and M. J. Van der Laan (2011): “Super learning,” in *Targeted Learning*, Springer, 43–66.

Rauch, J. and P. Evans (2000): “Bureaucratic structure and bureaucratic performance in less developed countries,” *Journal of Public Economics*, 75, 49–71.

Reinikka, R. and J. Svensson (2005): “Fighting Corruption to Improve Schooling: Evidence from a Newspaper Campaign in Uganda,” *Journal of the European Economic Associarion*, 3, 259–267.

Robinson, J. and T. Verdier (2013): “The Political Economy of Clientelism,” *Scandinavian Journal of Economics*, 115, 260–291.

Rose-Ackerman, S. and B. J. Palifka (2016): *Corruption and Government: Causes, Consequences, and Reform*, Cambridge university press.

Sachs, J. and A. Warner (1995): “The Big Rush, Natural Resource Booms, and Growth,” *Journal of Development Economics*, 59, 43–76.

Sequeira, S. (2012): “Chapter 6 advances in measuring corruption in the field,” in *New advances in experimental research on corruption*, Emerald Group Publishing Limited, 145–175.

Shaxson, N. (2007): “Oil, corruption and the resource curse,” *International Affairs*, 83, 1123–1140.

Svensson, J. (2005): “Eight Questions about Corruption,” *Journal of Economic Perspectives*, 19, 19–42.

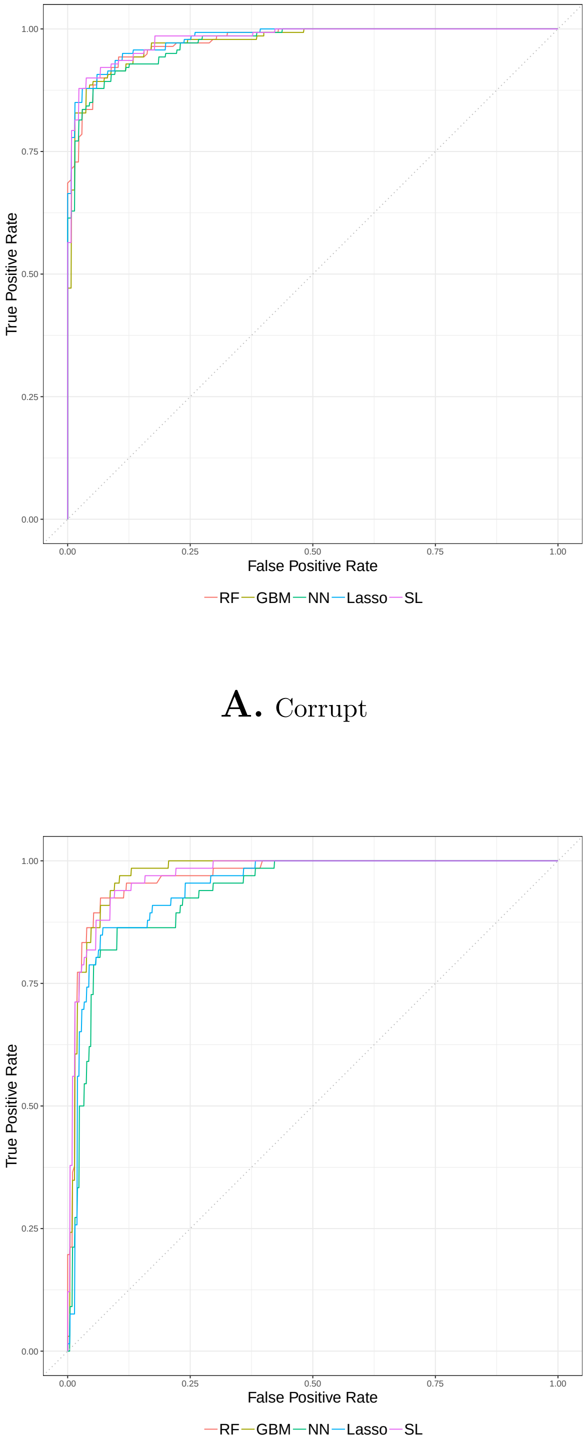
Tibshirani, R. (1996): “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007): “Super Learner,” *Statistical Applications in Genetics and Molecular Biology*, 6.

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, 307–331.

Xu, G. (2018): “The Costs of Patronage: Evidence from the British Empire,” *American Economic Review*, 108, 3170–98.

Figure 1. ROC curve

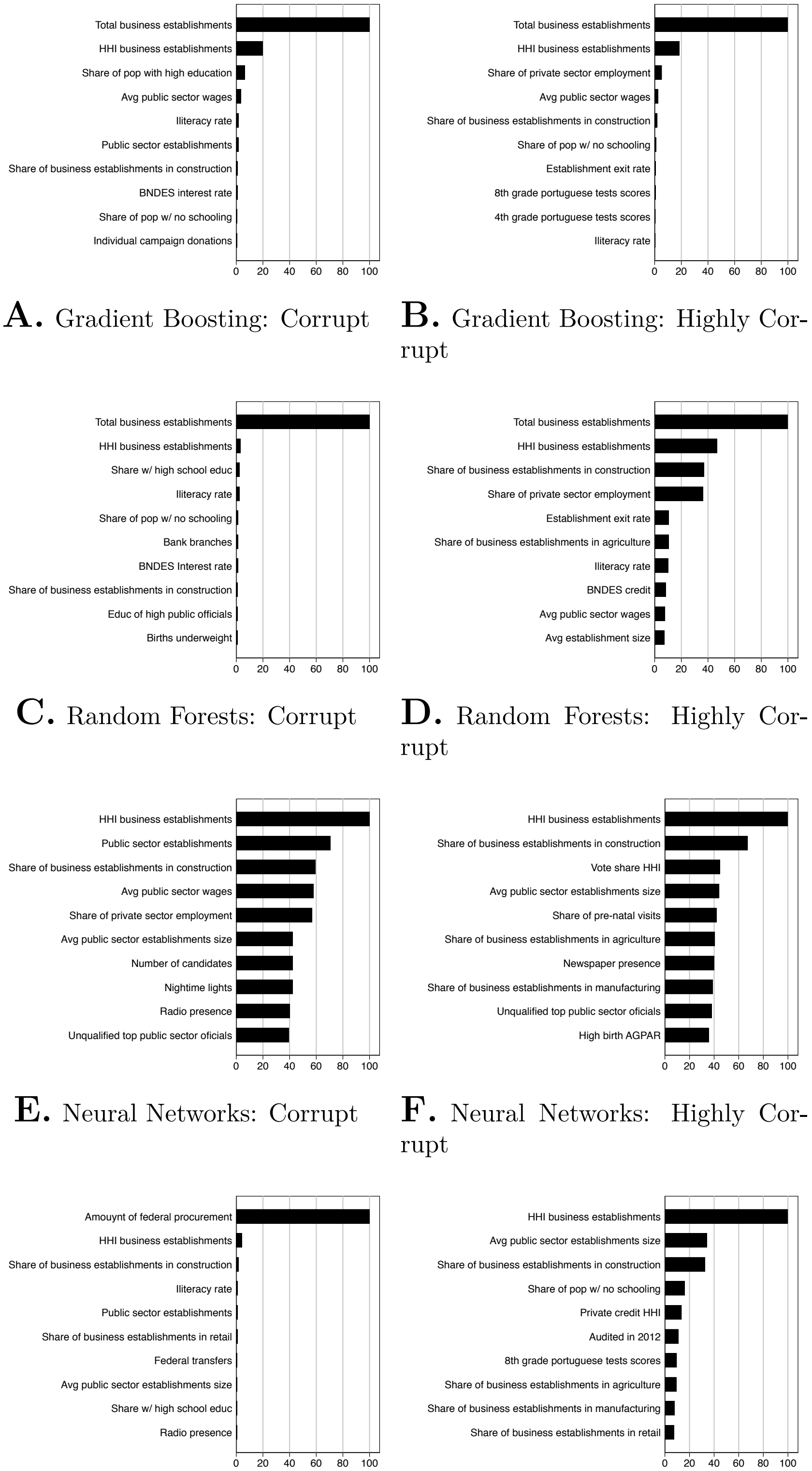


A. Corrupt

B. Highly Corrupt

Notes: This figure presents the ROC curves for all our models.

Figure 2. Covariates importance



A. Gradient Boosting: Corrupt **B.** Gradient Boosting: Highly Corrupt

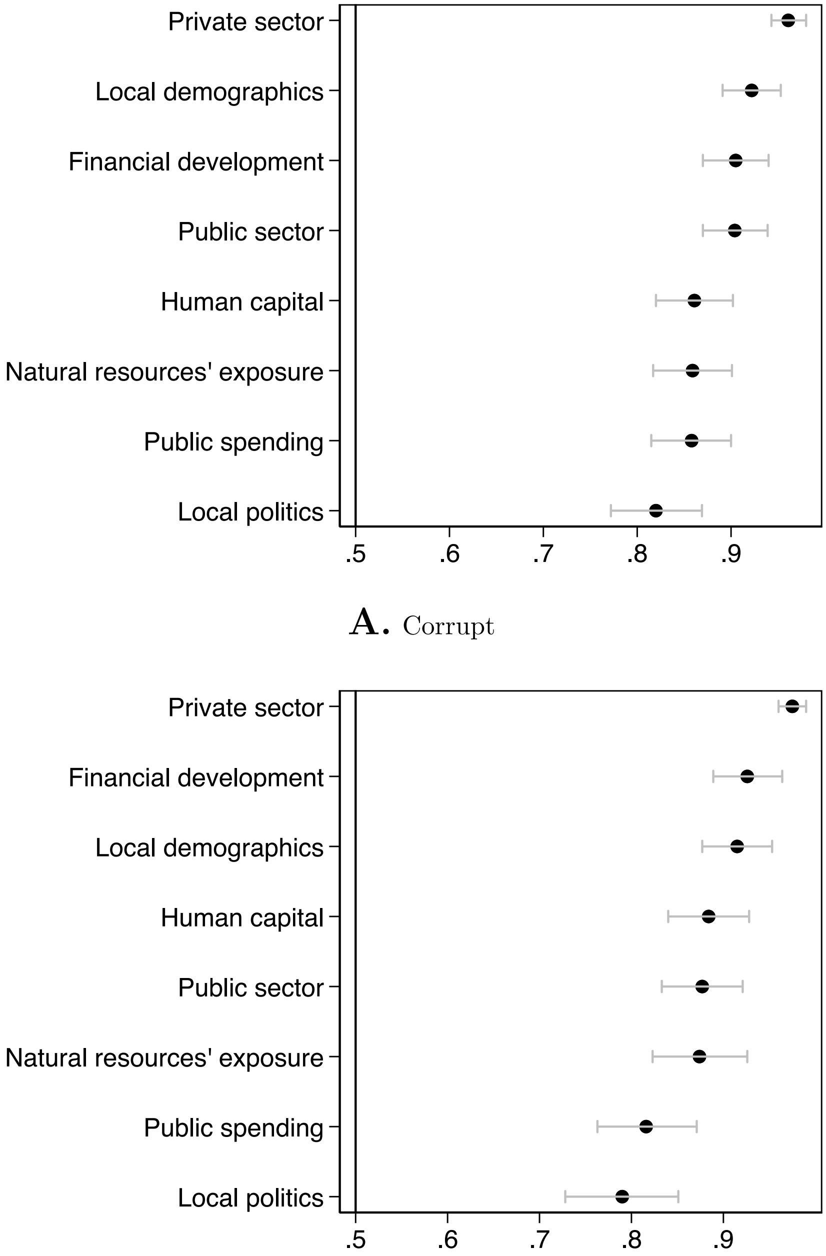
C. Random Forests: Corrupt **D.** Random Forests: Highly Corrupt

E. Neural Networks: Corrupt **F.** Neural Networks: Highly Corrupt

G. LASSO: Corrupt **H.** LASSO: Highly Corrupt

Notes: This figure presents the relative importance of covariates, as described in 5.4.

FIGURE 3. Group importance



A. Corrupt

B. Highly Corrupt

Notes: This figure presents the relative importance of group of covariates, as described in section 5.4. Confidence intervals at 95% are constructed by bootstrapping over the test set.

Table 1. Description of Variables

Categories	Source	Variables
Private sector	RAIS	Average business establishments size based on employment, number of business establishments, payroll per employee, average business establishments payroll, share of business establishments entering, share of business establishments exiting, business establishments churning, share of private sector workers over population, Hirschman-Herfindahl index based on business establishments size, average growth in business establishments and in employment in past 3 years, share of business establishments below 5 employees, share of business establishments between 5 and 25 employees, share of business establishments above 25 employees, share of business establishments in construction, share of business establishments in retail, share of business establishments in services.
Public sector	RAIS	Share of public sector employees over population, average wage of public sector employees, share of public institutions opening, share of public institutions closing, public institutions churning, share of workers by position within the institution, average growth in public employment and public institutions in past 3 years, share of public sector employees from municipal institutions, number of public institutions, average public institution size based on employment.
Financial development	BNDES ESTBAN, UNICAD	Share of business establishments receiving public loans, number of public loans per business establishment, total public credit per business establishment, average interest rate in public lending, bank branches per capita, banks per capita, total private credit per capita, total deposits per capita, and Hirschman-Herfindahl index based on private banks total assets and based on private banks credit.
Human capital	2000 Census, Ministry of Education, RAIS	Literacy rate, the share of population between 15 and 24 years old that finished, the first, second, and third cycle of primary education (Census), illiteracy rate (Census), average test scores in Portuguese and maths for nationwide tests at 4th and 8th grade, average private sector employees education, average private sector employees education by worker position within the firm, share of unqualified public employees based on job requirements, share of unqualified public employees by position within the institution, average public employees education, average public employees education by position within the institution, number of higher public education institutions per capita, number of higher private education institutions per capita.
Local politics	TSE	Number of candidates, Hirschman-Herfindahl index based on the vote shares, margin of victory between the winner and the runner-up, an indicator for whether the mayor is in his second term, an indicator for whether the mayor's party is the same as the one of the governor, an indicator for whether the mayor's party is from the same party as the one of the president, an indicator if the mayor is from right-wing party, an indicator if the mayor is from left-wing party, average candidate campaign donations and expenditures for firms and individuals, and per capita campaign donations and expenditures for firms and individuals.
Public spending	Ministry of Planning, Budget, and Management	Total expenditures per capita, personnel expenditures per capita, budget surplus per capita, total revenue per capita, federal transfers of capital per capita, federal current transfers per capita, transfers from the national tax fund per capita, share of business establishments in the municipality with public procurement, number of contracts per business establishments, federal procurement expenditure over population, share of discretionary contracts, and share of competitive contracts.
Local demographics	2000 Census, NOAA, Ministry of Health	Population density, GDP per capita, share of population living in rural areas (Census), deaths by aggression, GINI coefficient for income distribution (Census), average night light intensity coverage performing deblurring, inter-calibration, and geometric corrections, local radio, local newspapers, infant mortality rate, child mortality rate, average number of prenatal visits, share of abnormal births, share of underweight births, share of births with more than seven prenatal visits, and share of births with more than four prenatal visits.
Natural resources' exposure	RAIS IBGE	Share of business establishments in agriculture and mining sector, share of production of each of the top-7 crops in the country multiplied by the the log change in international prices and share of value of production over GDP (as constructed in Bernstein et al., 2018). The crops included are sugar cane, oranges, soybeans, maize, rice, rice, banana, and wheat, covering more than 98% of total agricultural production.

Table 2. Model's Parameters

	Optimal Parameters	
Model	Corrupt	Highly Corrupt
Lasso	λ: 0.01	λ: 0.01
Random Forest	Trees: 500	Trees: 500
	Mtry: 145	Mtry: 24
Gradient Boosting	Trees: 50	Trees: 50
	Depth: 1	Depth: 1
	Shrinkage: 0.1	Shrinkage: 0.1
	Min obs: 10	Min obs: 10
Neural Networks	Size: 5	Size: 5
	Decay: 0.1	Decay: 0.1
Ensemble Weights	Lasso: 0.05	Lasso: 0.08
	Random Forest: 0.22	Random Forest: 0.32
	Gradient Boosting: 0.55	Gradient Boosting: 0.60
	Neural Networks: 0.18	Neural Networks: 0

Notes: This table presents the optimal parameters for each of the prediction models we implement after the training procedure described in section 5.2.

Table 3. Model Performance

Model	LASSO	Random Forest	Gradient Boosting	Neural Networks	Ensemble
			Panel A: Corrupt		
AUC	0.97	0.97	0.98	0.95	0.98
Accuracy	0.91	0.91	0.92	0.88	0.92
Precision	0.91	0.93	0.92	0.89	0.94
Recall	0.92	0.89	0.93	0.89	0.91
F1	0.91	0.91	0.92	0.89	0.92
			Panel B: Highly Corrupt		
AUC	0.96	0.98	0.99	0.94	0.98
Accuracy	0.91	0.94	0.94	0.90	0.94
Precision	0.80	0.88	0.86	0.79	0.89
Recall	0.82	0.88	0.90	0.82	0.85
F1	0.81	0.88	0.88	0.80	0.87

Notes: This table presents the model performance for all our prediction models. *AUC*, *accuracy*, *precision*, *recall*, and *F1* are defined in section 5.3.

Table 4. Model Performance for Continuous Outcomes

Model	Baseline	LASSO	Random Forest	Gradient Boosting	Neural Networks
RMSE	8.08	6.39	4.94	4.77	8.37
MAE	4.37	3.13	1.80	1.82	2.79
R^2	0.00	NA	0.64	0.63	0.13

Notes: This table presents the model performance using the share of cases over establishments. *Baseline* model is the case in which the mean of the outcome is used as the prediction. *RMSE* is the root mean square error in the testing set, or the sample standard deviation of the differences between predicted values and observed values. *MAE* is the mean absolute error in the testing set, or the sample absolute difference between predicted values and observed values. R^2 is the in sample R-squared of the model.

Table 5. Model Performance for High Corruption Accounting for Class Imbalance

Model	LASSO	Random Forest	Gradient Boosting	Neural Networks	Ensemble
			Panel A: Over-sampling		
Accuracy	0.90	0.96	0.94	0.93	0.96
Precision	0.89	0.94	0.92	0.92	0.94
Recall	0.92	0.99	0.98	0.96	0.99
F1	0.91	0.97	0.95	0.94	0.96
AUC	0.96	0.99	0.99	0.97	0.99
			Panel B: Under-sampling		
Accuracy	0.87	0.91	0.96	0.86	0.94
Precision	0.87	0.93	0.95	0.87	0.95
Recall	0.89	0.91	0.97	0.88	0.93
F1	0.88	0.92	0.96	0.87	0.94
AUC	0.96	0.98	0.98	0.96	0.98

Notes: This table presents the model performance for the “Highly Corrupt” dummy accounting for class imbalance. In panel A, we perform over-sampling, in which observations of the minority class (highly-corrupt municipalities) are randomly replicated. In panel B, we perform under-sampling, in which observations of the majority class (non-highly-corrupt municipalities) are randomly excluded. *AUC*, *accuracy*, *precision*, and *F1* are as defined in section 5.3.

Table 6. Results from a doubly-robust LASSO

	Corrupt	Highly Corrupt	Share of Corrupt Cases
Employment HHI	0.326*** (0.013)	0.265*** (0.014)	3.549*** (0.398)
Sh private employees over population	-0.032*** (0.011)		
Sh of establishments in retail sector	-0.055*** (0.012)		
Sh rural population	0.035** (0.014)		
Local radio	-0.030*** (0.011)		
Number of candidates	-0.021* (0.012)		
Sh of establishments in construction sector		0.086*** (0.015)	1.723*** (0.391)
Sh of establishments in service sector		0.042*** (0.010)	0.586** (0.273)
Private credit HHI		-0.059*** (0.009)	
Sh of establishments in mining and agriculture		0.046*** (0.013)	
Sh of medium size establishment			1.063*** (0.366)
Sh of pop with more than 8 years of schooling			-0.334 (0.229)
Mean DV	0.508	0.255	3.836

Notes: This table presents the results for doubly-robust LASSO model suggested by Belloni et al. (2014).