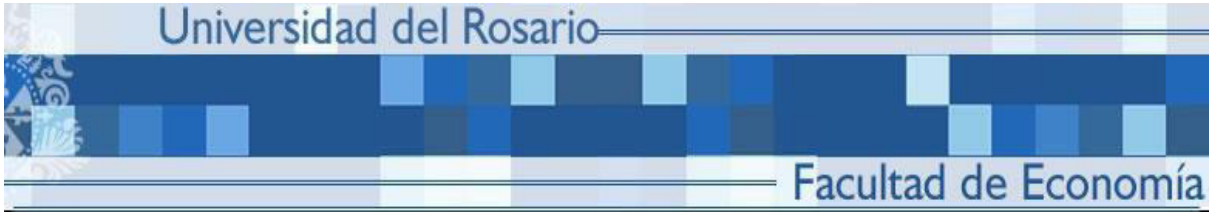


What Predicts Corruption?

Emanuele Colonnelli
Jorge Gallego
Mounu Prem

SERIE DOCUMENTOS DE TRABAJO

No. 228

Febrero de 2019

WHAT PREDICTS CORRUPTION?

EMANUELE COLONNELLI*, JORGE GALLEGO†, AND MOUNU PREM‡

Abstract. 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. 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.

Date: October 4, 2019.

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.

*Corresponding author: 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 presence of corruption, regulations tend to be inefficient (Djankov et al., 2002), businesses are held back (Fisman and Svensson, 2007; Colonnelli and Prem, 2019), 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. Using rich micro-data over the 2000-2014 period, we train multiple machine learning models to pin down what local characteristics best predict the intensity of corruption.

Our analysis reveals two important findings. First, machine learning models exhibit high levels of performance. That is, we can accurately predict where corruption is most likely to occur. Second, 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

[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.

relevant predictors, while variables related to the size and composition of the public sector, local politics, public spending, and natural resources' exposure have low predictive power.[3]

These findings have immediate implications. 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 difficult 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. Empirical Design

2.1. **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, which consists of randomized audits of municipal expenditures and subsequent public disclosure of investigative reports detailing all cases of irregularity and corruption uncovered. 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. Extensive details on the program are discussed by Avis et al. (2018). 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. We focus on two binary definitions of corrupt

[3]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 proven useful in other policy-related prediction issues (Kleinberg et al., 2015), such as security (Bogomolov et al., 2014), poverty (Blumenstock et al., 2015), and conflict (Blair et al., 2017; Bazzi et al., 2018). Lopez-Iturriaga and Sanz (2017) and Gallego et al. (2018b) study corruption using aggregate data and newspaper evidence from Spanish provinces and public procurement in Colombia, respectively.

municipalities, constructed using the share of total number of irregularities over the size of the municipality.[4] "Corrupt" ("Highly Corrupt") municipalities are those with an above median (top quartile) share in the distribution of corruption across all municipalities audited.

2.2. **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: private sector, public sector, financial development, human capital, local politics, public spending, natural resources' exposure, and local demographics. The data sources and exact definitions of each variable are reported in Table A1.[5]

2.3. **Machine Learning 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).

To train our models we divide the dataset into a training (70%) and a testing set (30%) and perform a 5-fold cross-validation to choose the optimal combination of parameters. In order to asses the models' performance we compute different measures in our test set. In particular, we compute the models' area under the ROC (Receiver Operating Characteristic) curve, namely the AUC, as well as each model's accuracy, precision, recall, and F1.[6] Extensive details on the estimation are reported in Appendix A.

[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.

[5]All variables, except the few in the Decennial Census, are measured as averages in the three years prior to the audit.

[6]The ROC curve plots the true positive rate versus the false positive rate, to measure the performance of a classifier. Such performance is usually summarized through the Area Under the Curve (AUC), with values close to 1 representing better classifiers. The *accuracy* of a model is the proportion of cases correctly classified. The *precision* is the proportion of positive *classifications* that are correct. The *recall* is the proportion of *actual* positives that are identified correctly. The *F1* score is a measure of the balance between precision and recall.

3. Findings

In this section we present the results of our analysis, focusing first on the overall performance of the predictive models and their robustness to alternative measures and specifications, and then on identifying the best individual and group predictors and their link to the corruption literature.

3.1. **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 1 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 1. Not surprisingly, the ensemble model performs best, as it is constructed by optimizing the weights of each individual model.[7]

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, 2019), but 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.

[7]In Appendix A.5 we present the robustness of our results to using a continuous measure of corruption and to account for class imbalance in the case of "Highly Corrupt."

3.2. **What Are the Best Predictors of Corruption?** We now move to the analysis of the individual covariates that best predict corruption. As standard in the literature, we focus on the tree-based models, and specifically on Gradient Boosting, as they allow us to quantify the information gain achieved when each predictor is used to partition the objects that are being classified (in our case, each municipality).

Panels C and D of Figure 1 plot 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.[8] The results highlight the striking importance of the 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. Figure A1 shows that these results are similar when estimating other machine learning models.[9]

Motivated by these individual ranking analysis, in Panel E and F of Figure 1 we perform an estimation where we categorize all 147 covariates into eight groups, as shown in Table A1. Sequentially and separately adding each group to the estimation of the ensemble model, we asses 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 F Figure 1).[10] The categories of *public sector*, *natural resources' exposure*, and *public spending* are less important predictors, and *local politics* is the least important one. 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 the size of the public sector to be strongly

[8]The importance of each covariate is standardized with respect to the 100-value of the top predictor.
[9]Similar patterns emerge by using the doubly-robust LASSO procedure proposed by Belloni et al. (2014) (see Appendix A.5).
[10]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.

linked to corruption (Robinson and Verdier, 2013; Gallego et al., 2018a; Colonnelli et al., 2019). 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). The emphasis on the role of the private and financial sector, on the other hand, remains significantly lower (Rose-Ackerman and Palifka, 2016).

4. 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 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.

References

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.

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," .

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.

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. and M. Prem (2019): "Corruption and Firms," Mimeo.

Colonnelli, E., M. Prem, and E. Teso (2019): "Patronage and Selection in Public Sector Organizations," Mimeo.

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.

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

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.

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 (2018a): "A Theory of Broker-Mediated Clientelism," Mimeo.

Gallego, J., G. Rivero, and J. Martinez (2018b): "Preventing Rather than Punishing: An Early Warning Model of Malfeasance in Public Procurement," Mimeo.

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.

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

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.

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.

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

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

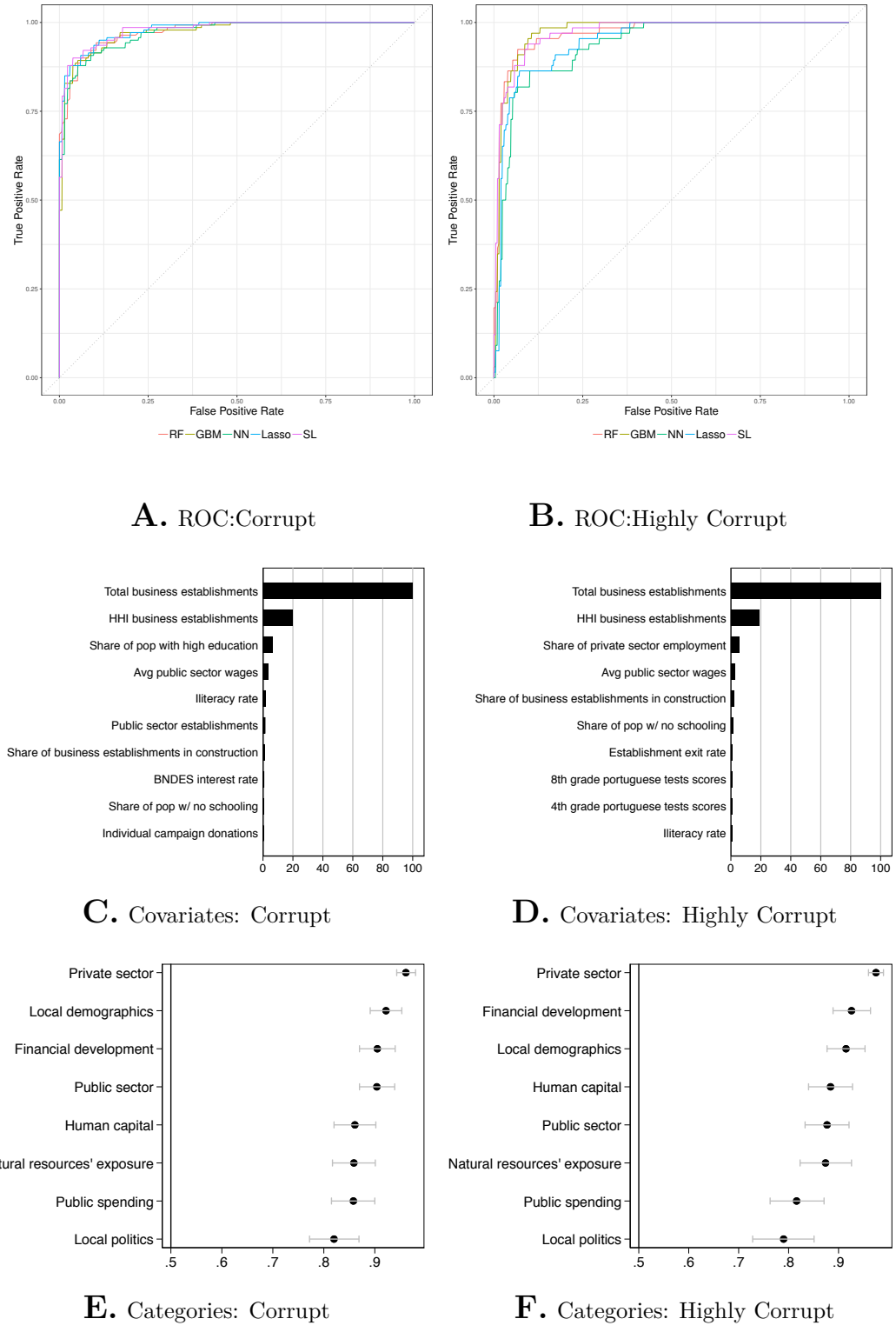
Rose-Ackerman, S. and B. J. Palifka (2016): *Corruption and Government: Causes, Consequences, and Reform*, Cambridge university press.

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.

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.

FIGURE 1. Predicting Corruption



A. ROC:Corrupt **B.** ROC:Highly Corrupt

C. Covariates: Corrupt **D.** Covariates: Highly Corrupt

E. Categories: Corrupt **F.** Categories: Highly Corrupt

Notes: This figure presents the ROC curves (panel A and B), which captures the trade-off between the true positive rate versus the false positive rate as we vary the discrimination threshold, for each of our prediction models. Panels C to F present the relative importance of covariates and categories, as described in the text and in Appendix A. Confidence intervals at 95% are constructed by bootstrapping over the test set.

TABLE 1. 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 the main text and in Appendix A.

Online Appendix

Appendix A. Machine Learning Models

A.1. **Models.** We train a set of different popular machine learning models in order to predict corruption. Each model has weaknesses and strengths, but in order optimize performance we also use an ensemble model, which combines the predictive capabilities of all of our individual models. In the end, we let the data inform which model is the more appropriate for this application based on their out-of-sample performance.

A.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 to identify complex relationships between the predictors and our outcome variable, i.e. corruption. The tuning parameter in this model is the weight of the penalization term in the objective function (λ), which is optimized over a grid of potential values using cross-validation.

A.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.

A.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 fittted sequentially to the full dataset and observations are weighted by the error rates of previous trees in the forest, allowing subsequent predictors to learn from the mistakes of the previous ones. Therefore, later trees are fitted with a larger weight on

observations that previous trees found difficult to predict. Consequently, as opposed to Random Forests, observations are not selected via bootstraping, but as function of past errors. In this way, the addition of each tree offers a slight improvement of the model (Freund et al., 1999). In our models, we keep fixed the learning rate (shrinkage parameter) and the minimum number of observations in terminal nodes, and use cross-validation to determine the optimal number of trees and the interaction depth.

A.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 use cross-validation to determine the optimal number of units in the hidden layer (size) and the regularization parameter (decay).

A.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.

A.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. 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) 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. Then, for each subsample, a model is fit in the other four subsamples and then test it in the remaining one. We repeat this procedure for each of the five subsamples and for each of value 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.
(4) Using these optimal parameters we fit our models in the testing set.

We standardize the data by the mean and standard deviation of the training set. Table A2 shows the optimal parameters for each of our models.

A.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 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; 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.

A.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 starts to decrease for the remaining variables. For the LASSO model the importance is determined

by the estimated coefficients of the regression. In the case of neural networks, importance is determined by the weights that connect variables within the network.

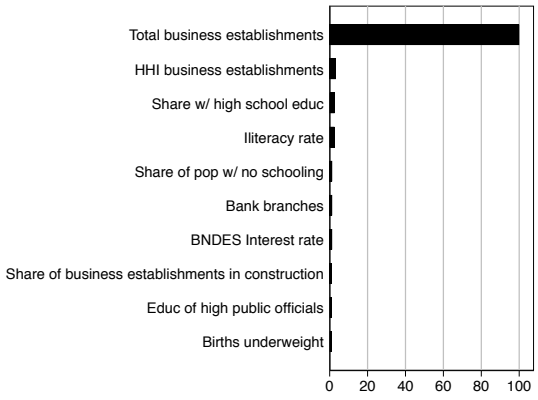
We then move to the analysis of the predictive performance of subgroups of related predictors in order to understand which categories matter the most. It may be the case that some groups do not have one particular variable that highly predicts corruption, but that the group as whole has a high predictive 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 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 level of predictive power. 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 exist any statistically significant differences in AUCs across categories.

A.5. **Robustness and Additional Analyses.** We estimate 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. 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 A3). Overall, our machine learning models perform better than the baseline case, with Random Forests and GBM usually achieving the highest levels of performance.

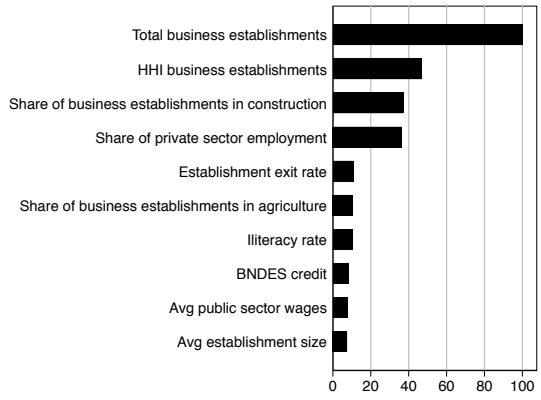
Additionally, we show that our findings for the "High Corrupt" dummy are robust to account for the class imbalance in the outcome. We use over- and an under-sampling techniques to randomly increase (decrease) the number of highly corrupt (non highly-corrupt) municipalities. Table A4 shows that our results remain largely unchanged.

Finally, we implement a variable selection procedure following Belloni et al. (2014). Table A5 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. In particular, these results show that private sector concentration (HHI) and the share of construction are positively correlated with corruption.

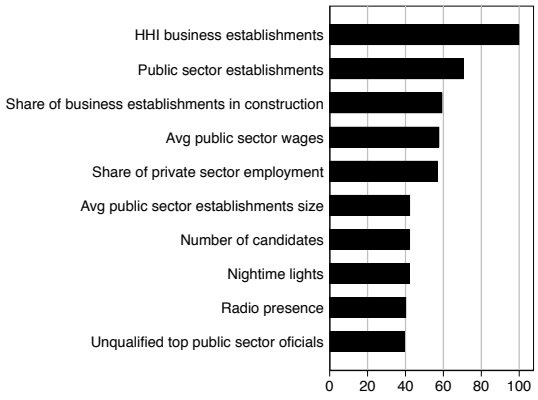
FIGURE A1. Covariates Importance for Other Models



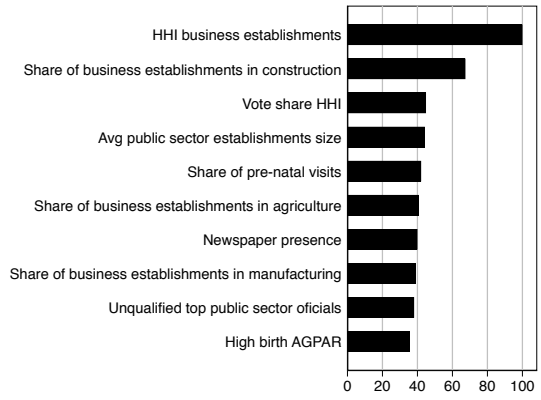
A. Random Forests: Corrupt



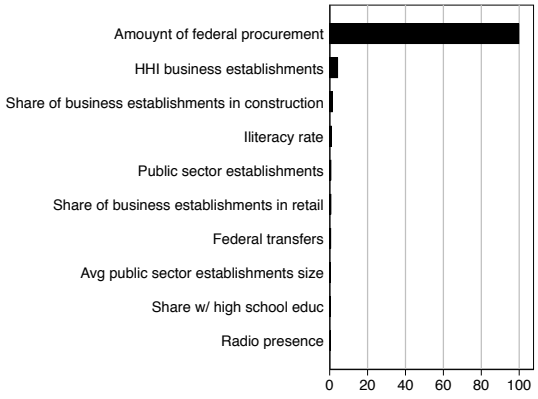
B. Random Forests: Highly Corrupt



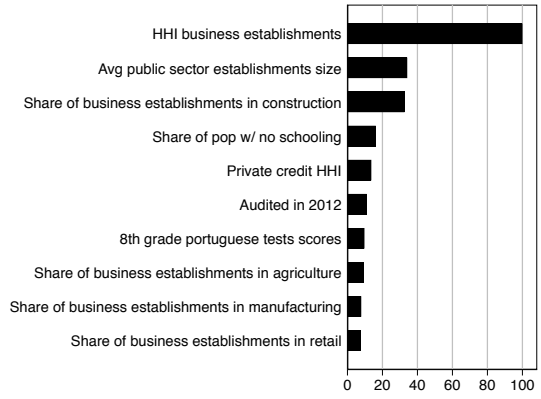
C. Neural Networks: Corrupt



D. Neural Networks: Highly Corrupt



E. LASSO: Corrupt



F. LASSO: Highly Corrupt

Notes: This figure presents the relative importance of covariates and categories, as described in the text and in Appendix A. Confidence intervals at 95% are constructed by bootstrapping over the test set.

Table A1. 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 nation wide 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 mayor's party is the same as the one of the governor, indicator for whether is from the same party as the one of the president, indicator if mayor is from right-wing party, indicator if 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, intercalibration, and geometric corrections, local radio, local newspapers, infant mortality rate, child mortality rate, average number of pre-natal visits, share of abnormal births, share of underweight births, share of births with more than 7 pre-natal visits, and share of births with more than 4 pre-natal 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 A2. Model's Parameters

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

Notes: This table presents the optimal parameters for each of the prediction models we implement after the training procedure. A brief description about each model can be found in Appendix A.

TABLE A3. 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 A4. 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 the main text and in Appendix A.

TABLE A5. Model Performance for High Corruption Accounting for Class Imbalance

	Corrupt	Highly Corrupt	Share of Corrupt Cases
Employment HHI	0.326***	0.265***	3.549***
	(0.013)	(0.014)	(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***	1.723***
		(0.015)	(0.391)
Sh of establishments in service sector		0.042***	0.586**
		(0.010)	(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).