OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 86, 1 (2024) 0305-9049
doi: 10.1111/obes.12586

Identifying Politically Connected Firms: A Machine Learning Approach*

VITEZSLAV TITL,†,‡,§ DENI MAZREKAJ§,¶,# and FRITZ SCHILTZ§

†*Utrecht University School of Economics, Utrecht University, Kriekenpitplein 21-22 Utrecht, 3584 EC, The Netherlands*
(e-mail: v.titl@uu.nl)
‡*Department of Economics, Faculty of Law, Charles University, Prague, Czechia*
§*Leuven Economics of Education Research (LEER), KU Leuven, Naamsestraat 69 Leuven, 3000*
(e-mail: d.mazrekaj@uu.nl; fritz.schiltz@kuleuven.be)
¶*Department of Sociology, Utrecht University, Padualaan 14 Utrecht, 3584 CH, The Netherlands*
#*Nuffield College, University of Oxford, New Road OX1 1NF, Oxford, UK*

Abstract

This article introduces machine learning techniques to identify politically connected firms. By assembling information from publicly available sources and the Orbis company database, we constructed a novel firm population dataset from Czechia in which various forms of political connections can be determined. The data about firms' connections are unique and comprehensive. They include political donations by the firm, having members of managerial boards who donated to a political party, and having members of boards who ran for political office. The results indicate that over 85% of firms with political connections can be accurately identified by the proposed algorithms. The model obtains this high accuracy by using only firm-level financial and industry indicators that are widely available in most countries. These findings suggest that machine learning algorithms could be used by public institutions to improve the identification of politically connected firms with potentially large conflicts of interest.

I. Introduction

In the heart of the second wave of the COVID-19 pandemic, on 26 November 2020, a controversial investigation was brought to light in a report published by the British

JEL Classification numbers: D72, D73, H83.

*The firm accounting data for this study are protected by a confidentiality agreement and we are precluded from sharing the data with others. Interested readers can consult the corresponding author for information on how to obtain access to the data. The code for all figures and tables is available at https://doi.org/10.5281/zenodo.10113144. We would like to thank Climent Quintana-Domeque for his guidance and valuable suggestions, Benny Geys, Kristof De Witte, Giovanna D'Inverno, Mark Verhagen, Lamar Pierce, and Aniek Sies for their useful comments and suggestions and also Alice Navratilova for excellent research assistance. Deni Mazrekaj acknowledges funding by the Research Foundation Flanders (FWO) (grant number 1257721N) and by the European Research Council (grant number 681546). Vitezslav Titl acknowledges support from the Horizon Europe project 'DemoTrans' (grant 101059288). The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

National Audit Office (2020). The spending watchdog found that more than half of the public pandemic contracts (£10.5 billion) related to personal protective equipment such as masks and protective gloves for healthcare workers, were awarded without a competitive tender. Nearly a third of these suppliers had links to politicians or senior officials and were referred to a 'high priority' channel, which was 10 times more likely to succeed in obtaining a contract than the regular competitive channel (Conn and Evans, 2020). Many of these suppliers had little or no experience in supplying personal protective equipment. For instance, a contract of £108 million was awarded to a chocolate wholesaler (Archer, 2020). In some cases, the paperwork stating why suppliers had been selected was missing and contracts were made only after the companies had already started the work (Pegg, Lawrence, and Conn, 2020).

Scandals involving links between politicians and private-sector firms (*political connections*) are by no means isolated incidents and can be found in virtually all countries. For instance, following a leak from the Panamanian law firm, Mossack Fonseca, the 'Panama Papers' revealed that the firm created thousands of shell companies for hundreds of politicians and public officials throughout the world (Harding, 2016). Evidently, not all entities involved in such political connections scandals are necessarily wrongdoers, but these examples highlight the need for transparency regarding political connections. This is especially the case given that the number of people and firms is persistently increasing, whereas budgets for audits are either remaining stagnant or are dropping. The United States Internal Revenue Service audited merely 0.45% of personal income tax returns in 2019, less than half of the audit rate in 2010 (Rubin, 2020).

In this article, we use supervised machine learning algorithms to predict political connections by constructing a novel firm population dataset from Czechia. Recently, machine learning algorithms have been found to improve predictions of many outcomes, such as poverty (Blumenstock, 2016; Jean *et al.*, 2016), teacher quality (Chalfin *et al.*, 2016), jail-or-release decisions (Kleinberg *et al.*, 2018), Post-Traumatic Stress Disorder (Abbasi, 2019) and even mortality (Puterman *et al.*, 2020). Ranking among the most corrupt countries in Europe according to Transparency International's Corruption Perception Index (Transparency International, 2019), Czechia is not a stranger to political connections scandals. On 4 June 2019 for instance, Czechia witnessed its biggest political protest since the fall of communism after the European Commission confirmed that Czech Prime Minister Andrej Babiš had significant conflicts of interest related to his private businesses. Specifically, his businesses received almost 20 million euros of EU agricultural subsidies while being Prime Minister (de Goeij and Santora, 2019). A unique feature of Czechia is that information on political connections is publicly available, although scattered. Many other countries such as France, Portugal, Canada, and the USA have introduced a ban on corporate donations to political parties, and information on firms' ownership structure and management is not available.[1] In Czechia, however, political donations are allowed, and firms' ownership structure and management can be retrieved. By employing

[1]Although banning corporate donations may appear as an effective policy to curb political connection at first sight, firms can still obtain connections by having their top officers (CEO, president, chairperson) affiliated with politicians or by politicians having equity in the firm (Faccio, 2006). These political connections are often even more difficult to track than corporate donations, leading to even less transparency than before the ban.

web scrapers and matching algorithms, we brought this information together, allowing us to observe political connections for the entire population of Czech firms. We consider firms as politically connected when they either have donated to a political party, have members of managerial boards who donated to a political party, or have members of (supervisory) boards who ran for office in the parliament, the Senate, a regional council, or a municipal council. Approximately 0.1% of the adult Czech population donates to a political party yearly.[2] This is in contrast with the situation in the USA, where about 9% of the population report having donated to a political party within the past year (Hughes, 2017). Our metric of political connections thus measures a rather unusual relationship between a firm and a political party. The overall share of firms with political connections in our sample is about 5%, which is similar to the share of politically connected firms in the USA according to Faccio (2006).

Politically connected firms generate substantial economic and welfare costs for society. For instance, Khwaja and Mian (2005) estimated that the costs of political connections may reach up to 1.9% of the GDP every year in Pakistan. These costs include higher product prices, poorly executed public works, hiring of less-competent individuals, erosion in employment standards, and an overall lack of efficiency (Cingano and Pinotti, 2013; Fisman and Wang, 2015; Titl and Geys, 2019; Colonnelli, Prem, and Teso, 2020; Titl, Geys, and De Witte, 2021; Baranek and Titl, 2020). Despite these negative implications of political connections, both firms and politicians have an incentive to become politically connected (Faccio, 2006; Sukhtankar, 2012; Cingano and Pinotti, 2013). Firms may benefit from politically channelled loans and contracts as well as regulatory benefits. On the other hand, politicians may garner votes and extract resources for political campaigns as long as the political connections remain unrecognized by the public. Given the large negative costs of political connections, it is critical to identify which firms are politically connected.

Our paper is closely related to the recent literature on 'prediction policy problems' in general (Kleinberg *et al.*, 2015), and on 'predictive policing' in particular, the idea that criminal activities can be predicted and therefore prevented before they happen (Brayne, 2017; Meijer and Wessels, 2019). For instance, Wheeler and Steenbeek (2020) use machine learning algorithms to predict robberies in Dallas (US), whereas Kondo *et al.* (2019) use them to detect and forecast accounting fraud. These types of algorithms seem to be very effective in combatting crime. Mohler *et al.* (2015) found that a machine learning algorithm used in the USA and the UK predicted 1.4–2.2 times more crime compared to a dedicated crime analyst. Similarly, Mastrobuoni (2020) estimated that 8 percentage points more robberies were solved as a result of predictive policing software used in Italy.

Machine learning algorithms have also been employed to predict corruption. In the absence of data on political connections, most studies were conducted at the aggregate level. Lima and Delen (2020) analyse cross-country data to predict and explain corruption across countries. Lopez-Iturriaga (2018) use the information on criminal cases involving a politician or a public official to estimate corruption risk for Spanish provinces. At a more local level, de Blasio, D'Ignazio, and Letta (2020) and Ash, Galletta, and Giommoni (2020)

[2]This is the average over the period from 2015 to 2019. All political parties in Czechia receive about 8,700 donations per year combined. The calculation is conducted by the authors based on data from PolitickeFinance.cz.

predict corruption crimes in Italian and Brazilian municipalities, respectively. Other studies have used more detailed, contract-level data, to detect corruption in public procurement in Colombia (Gallego, Rivero, and Martinez, 2021) and in Italy (Decarolis and Giorgiantonio, 2022). We contribute to this literature by constructing a novel dataset in which we measure actual political connections. Specifically, unlike previous studies, we are able to determine which firm is politically connected and which firm is not, going beyond country-, municipality-, and contract-level data. As such, we predict political connections by using machine learning algorithms at the level of actual political connections.

II. Data

Our data include a cross-section of all firms registered in Czechia in 2018.[3] Data on political donations were partly published in written reports held in the Parliamentary Library of the Czech Republic. We manually transcribed these reports into Microsoft Excel files. Another proportion of political donations was available on the website of the Office for Economic Supervision of Political Parties and Political Movements.[4] We merged these two sources of political donations and the combined dataset contains political donations to all political parties since 1995 from private citizens as well as firms. Our dataset ends in 2018 and contains in total 70,945 political donations. 59,600 donations were made by private citizens and the rest by firms. To obtain data on donating board members, we used a web scraper to download lists of board members of all Czech companies from the Czech company registry.[5] The company registry contains about 3,200,000 records. Each record represents an appointment term of a (supervisory) board member – their name, date of birth, academic title, city of residence and the beginning and end of the appointment. We matched the lists of individual persons who donated with the lists of (supervisory) board members of all Czech companies based on full name, date of birth, place of residence and academic title of each individual. Finally, the data on (supervisory) board members that ran for political offices were created by matching elections' candidate lists[6] and the lists of board members of all Czech companies mentioned above. The dataset of political candidates contains all candidates running for office in any municipal, regional, or central government (the Parliament) elections since 2004. It contains 740,000 records. Note that these are not distinct individuals as many politicians/political candidates have run repeatedly for office at different levels of the government and time. Part of the data on political donations and partially also personal connections are now available at the website PolitickeFinance.cz maintained by Datlab Institute.[7]

The data on predictors were obtained from the Orbis database collected by Bureau van Dijk. This database is available in most developed countries (about 450 million

[3]However, it is useful to test how our models generalize to other periods. Hence, in Figures S9–S11 in the online supplement, we also show the main results for the year 2011. The results are aligned with our findings for 2018 and therefore do not appear to depend on the year choice. We also show the feature importance ranking for boosting in 2011 in Figure S12. Three out of four variables overlap with our main analysis in 2018, with the difference that age has gained in importance, and the last year of submitted reports lost some of its importance.
[4]Accessible at https://www.udhpsh.cz/ (last accessed 14 November 2023).
[5]Accessible at https://portal.justice.cz/Justice2/Uvod/uvod.aspx (last accessed 14 November 2023).
[6]Accessible at https://www.volby.cz/ (last accessed 14 November 2023).
[7]Accessible at http://www.politickefinance.cz/ (last accessed 14 November 2023).

companies across the globe) and it provides standardized annual accounts (consolidated and unconsolidated), financial ratios, sectoral activities, and ownership data. We use all variables included in the Orbis database as predictors of political connections, except for categorical variables with extensively many categories as they could not be used by the machine learning models.[8] According to Czech law, all firms should submit their annual reports and yearly financial accounts to the company registry collected by Bureau van Dijk. Therefore, the Czech version of the dataset is more complete than datasets from other countries covered by the database such as the UK or Germany.[9] Nonetheless, the data do include some missing values, so we have included a value of 0 for missing data and we have added an indicator control for missing observations. Merging these financial data with our self-compiled political connections data was straightforward using company identifiers in both datasets. Lastly, we collected information about the value of public procurement contracts supplied by the firms and the value of subsidies from the European Union they received. This information is public in Czechia and was scraped from the official websites run by the Ministry of Regional Development.[10] The datasets were hand cleaned by a private company called Datlab, s.r.o.

The final dataset includes 254,367 firms, with each record containing financial and industry information as well as whether the firm was politically connected in 2018. We define political connections as an indicator given a value of 1 if the firm was politically connected and 0 otherwise. Firms are considered politically connected when they either have donated to a political party, have members of managerial boards who donated to a political party, or have members of (supervisory) boards who ran for office in the Czech parliament, the Senate, a regional council, or a municipal council. However, we also limit the definition of political connection to each of the three possibilities separately. Note that we do not observe, for instance, whether a firm is politically connected through a cousin or a best friend. We count 11,850 politically connected firms in 2018, comprising 4.65% of the overall sample. Descriptive statistics are presented in Table 1. Moreover, we observe in Figure 1 that politically connected firms are represented in all segments of the Czech market, regardless of the financial success of the firm.[11]

III. Methods

To predict political connections, we start with a logistic regression, which is widely used to predict binary outcomes. Then, we employ four commonly used supervised machine learning techniques: ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO), random forests, and random forests with boosting. All models have been performed using R version 3.6.3, and the script is available at https://doi.org/10.

[8]For instance, we do not use names of the auditors or city of headquarters as they constitute categorical variables with too many unique values.

[9]Note, however, that although the Czech coverage in Orbis is more complete than in other countries, most countries still do have a solid coverage of the main indicators, which allows for prediction of political countries in most countries covered in the Orbis database.

[10]Accessible at https://www.mmr.cz/cs/uvod and http://www.isvz.cz/ISVZ/Podpora/ISVZ.aspx (last accessed 14 November 2023).

[11]In Figures S13–S15, we show the same figure for the three largest sectors: retail (71,852 firms), real estate (34,369 firms), and manufacturing (28,182 firms), separately.

TABLE 1
Descriptive statistics

Predictor variable	*Mean*	*SD.*	*Min.*	*Max.*
Profit margin	0.7	16.5	−100	100
Return on capital	1.8	36.8	−984	1,000
Solvency ratio	20.6	37.7	−100	100
Number of employees[a]	12.0	93.3	0	10,000
Number of director managers[a]	1.5	1.1	0	55
Number of subsidiaries	0.02	0.4	0	113
Age (in years)	9.7	6.1	0	92
Total assets (mil. EUR)	1.4	43.3	−18	16,806
Operating revenue (mil. EUR)	1.6	28.2	−42	6,710
Profit and loss (mil. EUR)	0.1	4.4	−283	1,717
Profit before tax (thous. EUR)	64.4	5,146.0	−256,350	2,071,165
Cash flow (thous. EUR)	95.5	3,302.6	−187,811	686,970
Market capitalisation (mil. EUR)	0.1	36.3	0	17,336
Number of recorded shareholders	0.4	0.6	0	40
Shareholders' funds (thous. EUR)	638.6	23,507.1	−440,914	6,706,487
Financial expenses (thous. EUR)	73.2	4,594.2	−7,996	1,546,976
Operat. profit and loss (thous. EUR)	64.0	4,723.4	−256,752	2,002,007
Value of public procurement	37,424	3,592,569	0	1,244,676,100
Value of EU subsidies	17,240	904,408	0	279,395,085
Last year of submitted reports	2,015.54	1.515	2,005	2,018
Based in Prague	0.3	0.5	0	1
Politically connected	0.05	0.2	0	1

[a]Having zero directors/employees means that this firm is economically inactive. It was relatively costly in Czechia to dissolute companies, but cheap to keep them registered as inactive.

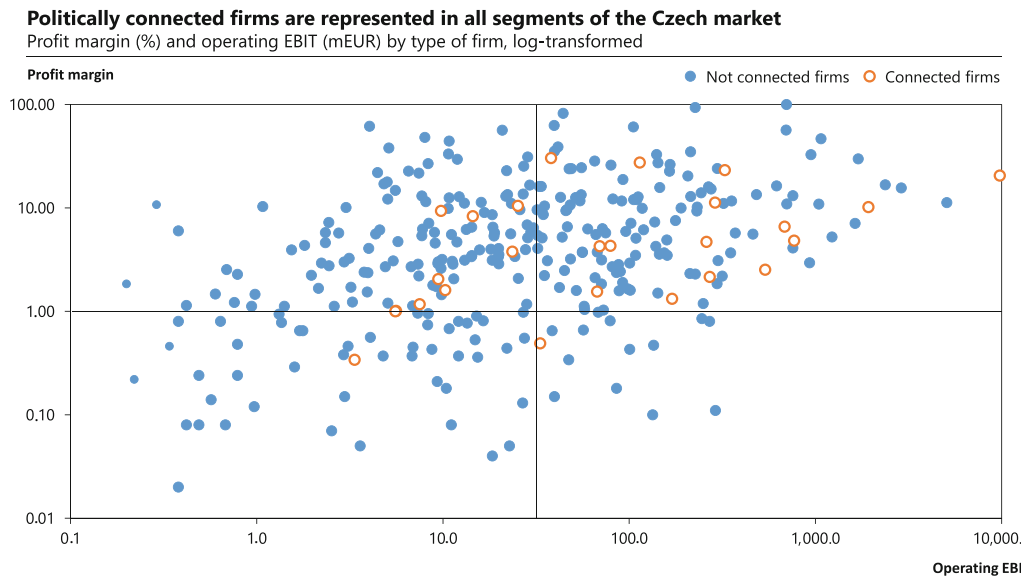

Figure 1. Politically connected and not connected firms according to firm profit margin and operating EBIT [Colour figure can be viewed at wileyonlinelibrary.com]

5281/zenodo.10113144. For each method, we divide the sample into a training set used to estimate the parameters of the models and a test set in which we predict political connections using the models. This is because using the same sample to both estimate the model and predict political connections leads to a training error rate that may dramatically underestimate the true error rate once the model is estimated on a different sample. In comparison to the training error rate, the test error rate is a better approximation of the true error rate (James *et al.*, 2013).

In our sample, only 4.65% of firms are politically connected. This is problematic because 95.35% of firms will be correctly identified when firms are always predicted not to be politically connected. To prevent the algorithms to achieve high accuracy by always predicting the most common group, we follow the literature on corruption prediction (de Blasio *et al.*, 2020) and use the synthetic minority oversampling technique (SMOTE) (Chawla *et al.*, 2002). This technique essentially randomly undersamples the majority class, that is, not politically connected firms. Instead of estimating the models on all politically connected and not-politically connected firms in the training set, we balance the number of politically connected and not-politically connected firms by randomly taking a subset of the not-politically connected firms. For instance, if our training set includes 5,000 politically connected firms, we randomly draw 5,000 not politically connected firms to be used in the training set.

Although randomly dividing the sample into a training and a test set leads to better predictions than solely using the training set, this random division can lead to a test error rate that can be highly variable depending on the observations that are included in the two sets. A commonly used approach to reduce this variability is the k-Fold Cross-Validation resampling method. This method randomly divides the set of observations into k non-overlapping groups (*folds*). For each group, the sample is divided into a training set and a test set, and the classification error rate is calculated. The classification error rate is the ratio of the number of firms that were incorrectly classified as politically connected and not politically connected over the total number of firms in the test set. The mean classification error rate is then computed by averaging the k classification rates obtained from the different folds. In our application, we opted for the commonly used 10-fold cross-validation in which data are split into a 90% training set and a 10% test set for each fold. This choice was made for three reasons. First, 10-fold cross-validation is widely used in the literature which aids in reproducibility and comparison with other studies. Second, it is computationally efficient as it only estimates the models 10 times. Lastly, it is beneficial to keep the training set large as models tend to be more efficient in large samples, reducing the variance of the test error rate.

We compare methods based on their accuracy of prediction: the number of correctly classified firms divided by the total number of firms. Further, we also estimate the sensitivity and specificity of each model. The sensitivity of a model is the number of correctly classified politically connected firms divided by the total number of correctly classified firms. Analogously, the specificity of a model is the number of correctly classified not politically connected firms divided by the total number of correctly classified firms. Calculating the sensitivity and specificity of the models is useful because it is more costly to believe that politically connected firms are not politically connected than vice-versa from a policy perspective.

Logistic regression

We start with a logistic regression model used for binary outcomes. It can be formulated as follows:

$$\log\left(\frac{P(Y_i = 1 \mid \boldsymbol{X_i})}{1 - P(Y_i = 1 \mid \boldsymbol{X_i})}\right) = \beta_0 + \delta \boldsymbol{X_i}, \quad (1)$$

where Y_i is indicator given a value of 1 if firm i is politically connected and 0 if firm i is not politically connected and $\boldsymbol{X_i}$ is a set of predictors. The left-hand side of equation (1) specifies the log odds of being politically connected. We convert these log odds into probabilities ranging from 0 to 1. As common in the literature on prediction, we define a firm to be politically connected if the probability of being politically connected exceeds 50%. This is a more conservative approach than the approach based on the Receiver Operating Characteristic (ROC) curve in which the researcher seeks a threshold to maximize the model performance. Nonetheless, we opted for the conventional 50% threshold because it is widely used, intuitive, and the model performs well regardless. Thus, our models estimate a lower bound and we use this conservative approach for all models that follow.

Shrinkage estimators

It is unlikely that all the predictors used in the logistic regression in equation (1) are useful in predicting political connections. Including irrelevant variables leads to unnecessary complexity, the risk of overfitting, a higher variance in prediction, and a larger test set error. For this purpose, we use two common shrinkage estimators: ridge regression and least absolute shrinkage and selection operator (LASSO). In this approach, we fit a logistic regression that includes all predictors while shrinking (*regularizing*) some of the coefficients towards zero. This approach has been found to improve the fit by greatly reducing the variance of predictions while slightly increasing the bias.

Ridge binomial regression (linear version introduced by Hoerl and Kennard, 1970) maximizes a penalized version of the log-likelihood. From the standard binomial log-likelihood, a *shrinkage penalty* of the following form $\lambda \sum_{j=1}^{p} \mid \beta_j^2 \mid /2$ is subtracted (β here represents the regression coefficients). The *penalty* tends to shrink the coefficients towards zero. The tuning parameter λ sets the level of shrinkage. If λ is zero, the ridge regression resorts to a standard logistic regression. The higher the λ, the more ridge regression coefficients will approach zero, but never reach zero. Ridge regression is very sensitive to the scaling of each predictor. Therefore, we apply ridge regression after standardizing the predictors.

The potential disadvantage of ridge regression is that it does not exclude any of the coefficients. Although coefficients are shrunk towards zero, they never reach zero. LASSO (formalized by Tibshirani, 1996) overcomes this disadvantage by maximizing the log-likelihood with the following shrinkage penalty $\lambda \sum_{j=1}^{p} \mid \beta_j \mid$.[12] With LASSO, coefficients

[12]We performed both ridge regression and LASSO using the *SL.glmnet* function in the *SuperLearner* package in R. We used the default option of the package, which chooses the optimal tuning parameter λ that minimizes the classification error from 100 different values of the parameter.

can be exactly zero. Therefore, LASSO will select some of the variables and discard others. In contrast, ridge regression always includes all the variables in the model. Depending on whether all variables are relevant or not, one method may outperform the other.

Tree-based methods

The main disadvantage of logistic regression and shrinkage estimators is that interactions and nonlinearities (e.g., higher-degree polynomials) need to be modelled explicitly. With many predictors, this process is cumbersome and largely arbitrary. By contrast, tree-based methods capture interactions and nonlinearities by construction (Mullainathan and Spiess, 2017; Basu *et al.*, 2018). The classification tree algorithm considers all possible splits of all predictors and chooses the one that minimizes classification error. The most predictive split (which reduces classification error the most) is placed on the top of the tree. Repeating this process from top to bottom results in the construction of a classification tree.

A limitation of classification trees is that they suffer from high variance. A small change in the training data can lead to a large change in the estimated tree. The accuracy of predictions can be improved when combining information from several classification trees into an ensemble method called 'random forest', pioneered by Ho (1995) and later Breiman (2001). In this algorithm, several random samples are drawn from the training set and a decision tree is grown on each sample (*bagging*). Moreover, a random subset of the predictors is chosen as possible split variables at each split. To aggregate trees, each tree is given one vote and firms are classified by a majority vote.[13]

Another possible improvement to classification trees is boosting, proposed by Friedman (2001). Unlike a random forest that constructs trees independent of the other trees, the boosting algorithm operates iteratively and constructs trees sequentially by learning from the previously constructed trees. As each tree is constructed using information from previously constructed trees, smaller trees are typically sufficient. The boosting algorithm learns sequentially by first growing a classification tree and then reweighting the data for the next classification tree. Misclassified observations get more weight.[14]

IV. Results

We predict political connections with logistic regression and four commonly used supervised machine learning techniques: ridge regression, least absolute shrinkage and selection operator (LASSO), random forests, and boosting. Figure 2 reports the prediction accuracy of different models on the test set, namely on a sample that the algorithm has not yet seen before. For instance, the figure shows that if 5,000 connected firms and 5,000 not connected firms are used to train the algorithm, boosting is 84.1% accurate in predicting which firms are politically connected on a subsample of the same size randomly drawn from the rest of the sample.

[13]We perform random forests using the SL.randomForest function in the SuperLearner package in R. In our case, we use the default options of the package: 1,000 trees are grown and the number of predictors used in each tree is set to the square root of the total number of predictors.

[14]We performed boosting using the SL.XGBoost function from the SuperLearner package in R. As the tuning parameters, we opted for the default values: the number of trees equals to 1,000, the maximum depth of a tree equals 4, and the minimum number of observations allowed per tree nodes equals 10.

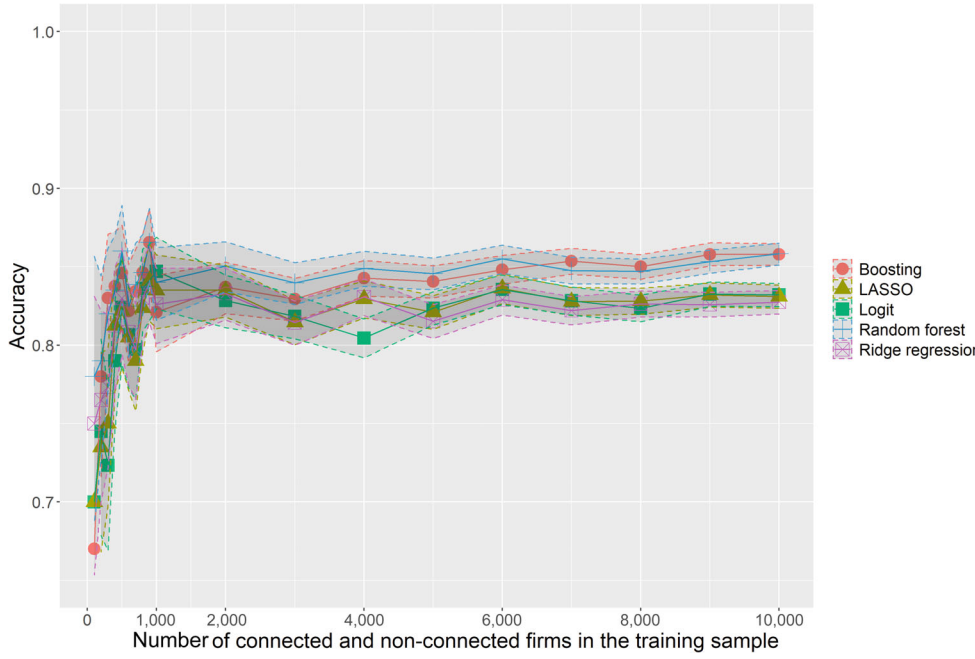

Figure 2. Accuracy of predicting political connections using machine learning.
Notes: The figure can be interpreted as follows: if 5,000 connected firms and 5,000 not connected firms are used to train the algorithm, boosting is 84.1% accurate in predicting which firms are politically connected on a subsample of the same size randomly drawn from the rest of the sample. The figure displays 95% confidence intervals [Colour figure can be viewed at *wileyonlinelibrary.com*]

Figure 2 shows that all five models are highly able to predict which firms are politically connected. It appears that random forests and boosting perform best, especially when the number of firms used to train the algorithm is large. Nonetheless, even with merely 200 firms, all algorithms predict political connections with about 75% accuracy, much higher than under random auditing. We further examine whether this high accuracy stems from the correct prediction of politically connected or not politically connected firms. From a policy perspective, it is more costly to believe that politically connected firms are not politically connected than vice versa. Figure 3 shows the true positive rate, namely the rate at which politically connected firms are correctly predicted (*sensitivity*). Figure 4 shows the true negative rate, the rate at which not politically connected firms are correctly predicted (*specificity*). Corresponding confusion matrices are provided in Table S2. Given that the true positive rate is mostly higher than the accuracy overall, it appears that the high accuracy mainly stems from correctly predicting politically connected firms. Especially boosting and random forests are better at predicting politically not connected firms, compared to the other methods. Random forest and boosting overperform other methods in all aspects (as visible partially also visible from Figures 2–4).

To gauge how the algorithms make their decisions and what kind of firms cultivate political connections, we provide coefficient estimates for logistic regression in Table S1 and a feature importance ranking for boosting in Figure S1. It appears that four factors – age, value of public procurement, operating revenue, and the last year of

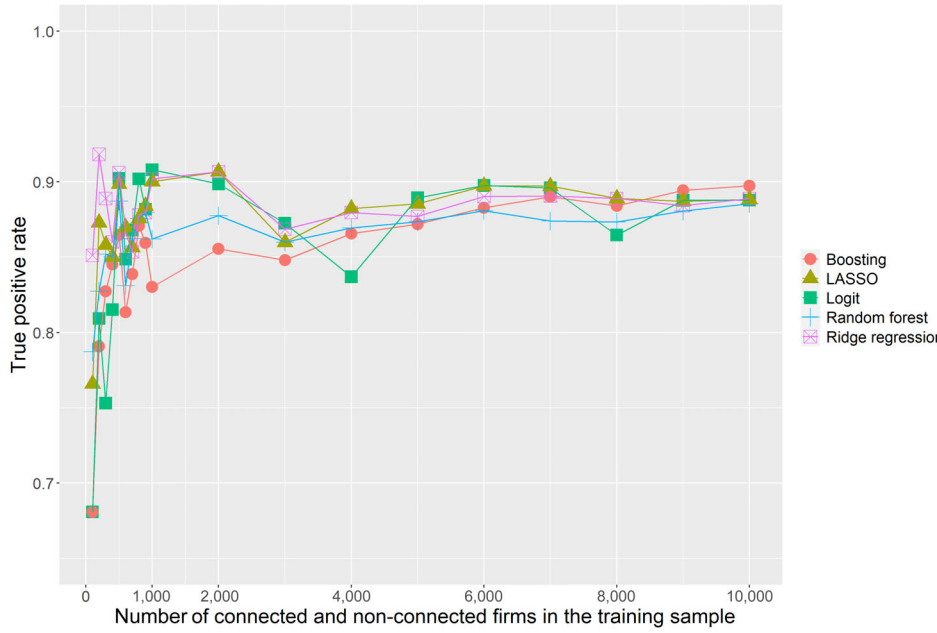

Figure 3. Sensitivity of predicting political connections using machine learning.
Note: The figure can be interpreted as follows: if 5,000 connected firms and 5,000 not connected firms are used to train the algorithm, boosting predicts 87.2% of politically connected firms correctly in a subsample of the same size randomly drawn from the rest of the sample [Colour figure can be viewed at *wileyonlinelibrary.com*]

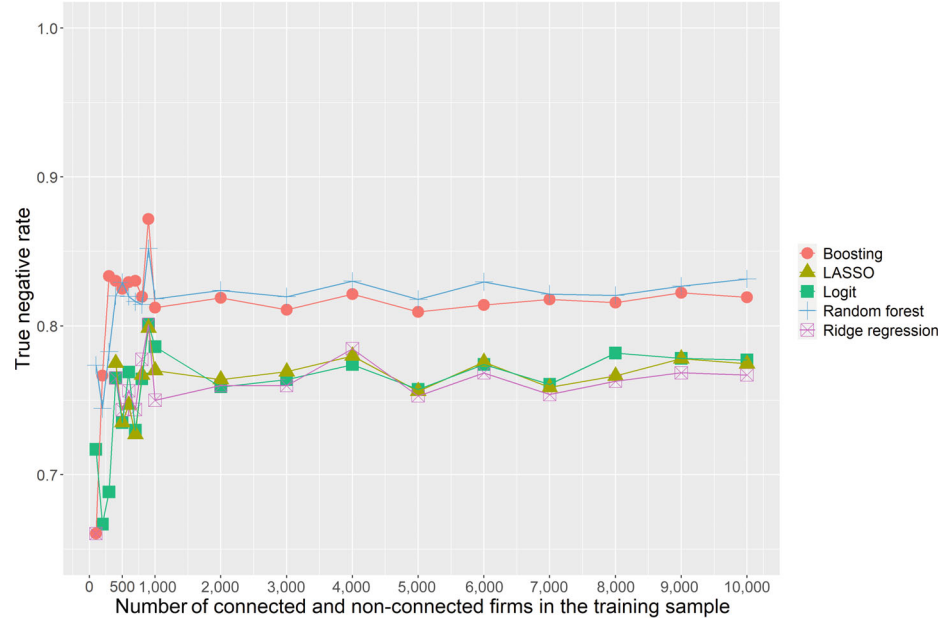

Figure 4. Specificity of predicting political connections using machine learning.
Note: The figure can be interpreted as follows: if 5,000 connected firms and 5,000 not connected firms are used to train the algorithm, boosting predicts 81% of not-politically connected firms correctly in a subsample of the same size randomly drawn from the rest of the sample [Colour figure can be viewed at *wileyonlinelibrary.com*]

submitted reports – have the largest predictive power for political connections. The older the company is, the more time it has to establish connections to politicians. Also, firms with connections enjoy various benefits such as access to loans (Claessens, Feijen, and Laeven, 2008) and preferential treatment in public procurement (Titl and Geys, 2019). This suggests that they are more likely to survive longer than firms without political connections. The literature on the preferential treatment of politically connected firms on public procurement markets also helps to explain why the value of public procurement contracts is an important predictor. Larger firms are also more likely to be connected according to the previous literature (Faccio, 2010) – this helps to explain why operating revenue is a strong predictor. Finally, the role of the last year of submitted reports variable might appear surprising, but it can be seen as a measure of compliance with the law that requires the publication of annual reports in the company registry. Ferwerda et al. (2017) show that such behaviour is a good predictor of corruption in public procurement. The order of importance of the predicting variables is likely a country-specific finding; and hence, in general it is necessary to consider a wider range of financial and industry indicators to predict political connections with a high accuracy.

Building on this insight, we also show the partial dependence plot of the two strongest predictors (age and the value of public procurement contracts) and the likelihood of being connected in Figure S2. The figure shows that the likelihood of being connected increases with the increase in age and with a decline in the value of procurement contracts. Similarly, we also calculated the relative importance of predictors in the random forest model using the mean decrease in Gini. This indicator measures how each variable contributes to the homogeneity of the tree nodes. Ranking the variables based on this indicator, the top five predictors are age, sector, operating revenue, profit before taxes, the last year of submitted reports, with Gini scores ranging from 72 to 14. We obtain a comparable picture when using the mean decrease in accuracy, with a correlation between the mean decrease in Gini and accuracy at 0.88. We conclude that there is no single variable that would accurately predict political connections, but rather a flexible combination of many variables is needed.

Until now, we defined a firm to be politically connected when they either have donated to a political party, have members of managerial boards who donated to a political party, or have members of (supervisory) boards who ran for office. However, this definition may be too inclusive and does not entail a two-way relationship which would normally be implied by 'connection'. For instance, personal donations of a board member do not necessarily reflect the intention of the firm to create a political connection. Hence, we estimate how the predictive performance changes when we limit the definition of political connection to each of these possibilities. Figures 5–7 present the accuracy rates of the same set of methods for the three measures of political connections separately. As visible from the figures, the accuracy rates are in general higher than for the definition encompassing all three measures. Also, the accuracy rate is again the highest for the boosting method. The accuracy rate (with the training and test set sample of 5,000 firms) is achieving 97.1% in the case of prediction of firms with their managerial boards' members donating to political parties, but it is above 90% for all three measures. In Figure S3-S8, we also present the sensitivity and the specificity rates for all three measure of political connections separately.

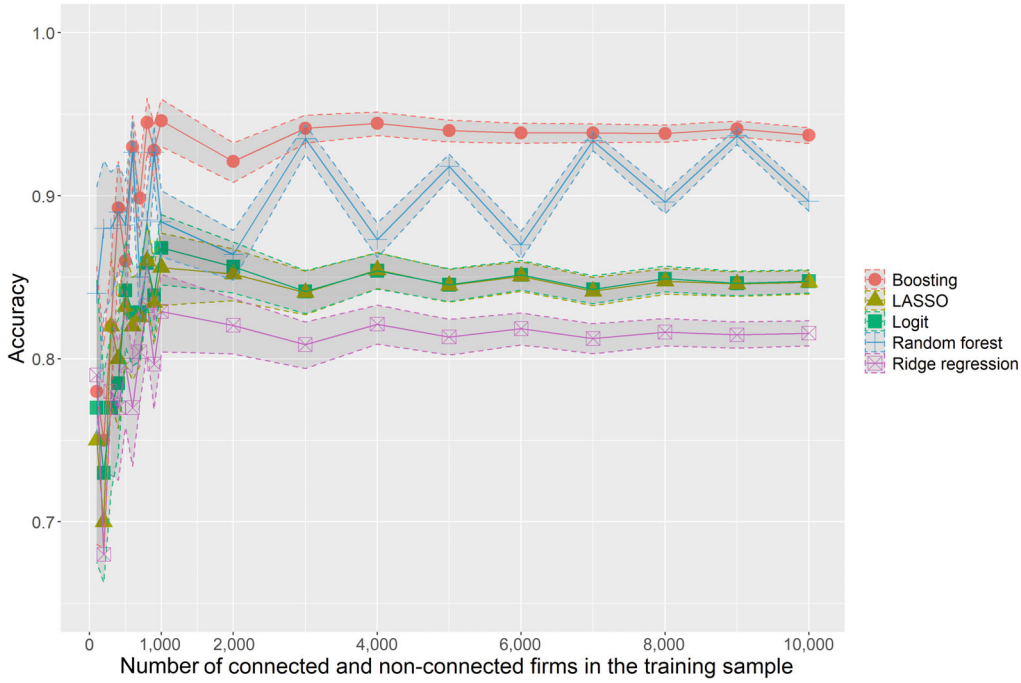

Figure 5. Accuracy of predicting political donors using machine learning.
Note: The figure can be interpreted as follows: if 5,000 donating firms and 5,000 not-donating firms are used to train the algorithm, boosting is 94% accurate in predicting which firms are politically connected on a subsample of the same size randomly drawn from the rest of the sample. The figure displays 95% confidence intervals [Colour figure can be viewed at *wileyonlinelibrary.com*]

V. Discussion

We used supervised machine learning algorithms to predict political connections by constructing a novel firm population dataset from Czechia. The models obtained high accuracy, higher than the conventional logistic regression, by using only firm-level financial and industry indicators that are widely available in most countries. These results suggest that machine learning algorithms could be used by public institutions to help identify firms whose political connections could represent major conflicts of interest. Firms identified by the algorithms as politically connected can be targeted for inspection. Thereby, we could avoid welfare losses stemming from politically connected firms (reaching up to 1.9% of GDP every year in Pakistan according to Khwaja and Mian, 2005). These losses include higher product prices, poorly executed public works, hiring of less-competent individuals, erosion in employment standards, and an overall lack of efficiency (Cingano and Pinotti, 2013; Fisman and Wang, 2015; Titl and Geys, 2019; Colonnelli *et al.*, 2020).

In this respect, the Ukranian system 'Dozorro' can be used as an inspiration (Observatory of Public Sector Innovation, 2016). This system employs machine learning algorithms in public procurement to detect tenders with a high level of corruption. Once the algorithm detects suspect tenders or purchases, they are reported to the authorities to be investigated. Given the high costs of political connections and the low share of

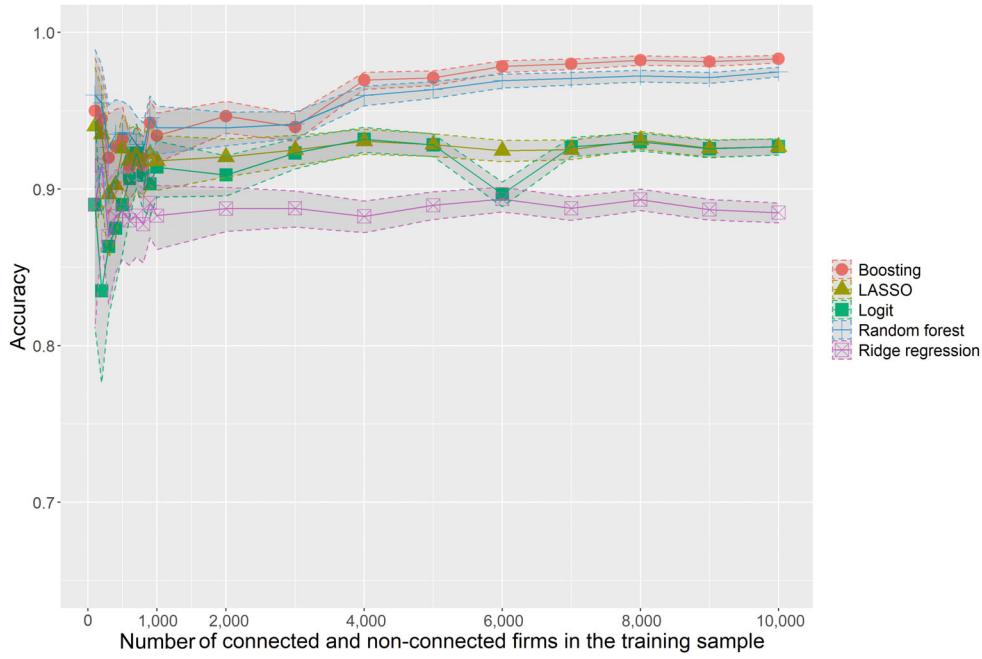

Figure 6. Accuracy of predicting CEO donors using machine learning.
Note: The figure can be interpreted as follows: if 5,000 firms with their CEOs donating and 5,000 firms with their not-donating CEOs are used to train the algorithm, boosting is 97.1% accurate in predicting which firms have CEOs that donate to political parties on a subsample of the same size randomly drawn from the rest of the sample. The figure displays 95% confidence intervals [Colour figure can be viewed at *wileyonlinelibrary.com*]

firms that are randomly inspected, targeted audits based on machine learning algorithms could deter firms from malpractice as they would have a higher chance of being inspected than under random auditing. Nonetheless, we believe that targeted audits would be most effective if also some of the randomness in inspections were maintained. Further studies should analyse the optimal ratio of targeted to random inspections and investigate whether targeted algorithmic inspections may have larger spillover effects on non-inspected firms than random inspections.

It is useful to reflect on how our findings could be relevant for countries other than Czechia. The case of Czechia was particularly suitable for this study because we could construct a dataset that includes a complete universe of three distinct types of measures for political connections. This information is not readily available in most other countries. To apply our algorithm, however, a public agency or an enforcement body does not need complete data on political connections. Only a small sample of a few hundred connected and non-connected firms and a set of predictors such as financials of firms from Orbis are necessary to train the algorithm. This makes the application in most developed countries feasible. A complete sample is only needed to assess the accuracy of such models, which was our goal. For example, Central Eastern European countries such as Slovakia and Poland as well as South-European Countries such as Italy and Spain have a rather good coverage of the necessary indicators and suffer from similar levels of corruption so the usage of an algorithm similar to ours could be beneficial in fighting adverse effects of political connections.

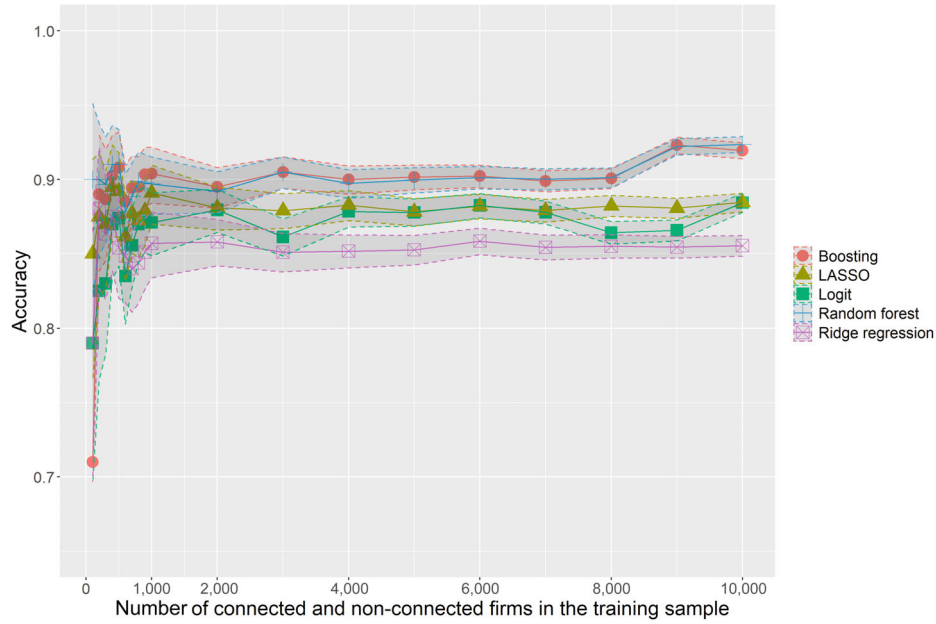

Figure 7. Accuracy of predicting personal political connections using machine learning.
Note: The figure can be interpreted as follows: if 5,000 personally connected firms and 5,000 not-connected firms are used to train the algorithm, boosting is 90.1% accurate in predicting which firms are personally politically connected on a subsample of the same size randomly drawn from the rest of the sample. The figure displays 95% confidence intervals [Colour figure can be viewed at *wileyonlinelibrary.com*]

The use of machine learning to increase the probability of detecting firms that are associated with the aforementioned welfare losses is likely to be beneficial for the society. There are, however, pitfalls to these methods. Such algorithms tend to amplify potential biases. In our case, this could be (discriminatory) focus on, for example, foreign owned firms. If a few companies with foreign ownership are involved in a political connections scandal and are used to train the algorithms, foreign owned firms may be more likely to get flagged than domestic firms, even though most foreign firms operate in a wholly legitimate manner. Furthermore, over time, we expect firms to become familiar with machine learning algorithms and then firms may improve their ability to fool the algorithm. This could be overcome by persistently updating the algorithm parameters for further targeted inspections (Ash *et al.*, 2020) and by random audits that would ensure that even low-risk firms have some chance of being inspected.

Although machine learning algorithms appear to predict political connections with great accuracy, the algorithms are not always easily interpretable. In this article, we have used relatively simple machine learning algorithms and have refrained from using black-box models such as neural networks. These black-box algorithms store nonlinear relationships between variables in a nonobvious form (Murdoch *et al.*, 2019), but in return achieve even greater predictive accuracy. It is therefore likely that even more politically connected firms could be predicted accurately if more complex algorithms were used at the expense of interpretability. Similarly, information on the demographic and other characteristics of the members of managerial and supervisory boards are not included in the dataset we have

access to. Although these characteristics would be very useful, it should be noted that even without these potential strong predictors, the algorithms achieve high accuracy rates, which is good news from policy perspective as such data may not be available in many countries. Thus, our results indicate that even relatively easily interpretable algorithms with a limited set of variables can predict political connections with high accuracy.

Final Manuscript Received: February 2022

References

Abbasi, J. (2019). 'First biomarker-based screening tool for PTSD', *Journal of the American Medical Association*, Vol. 322, p. 1437.

Archer, B. (2020, July 9). Legal proceedings against UK government over awarding of £108m PPE contracts to Antrim firm. Retrieved from The Irish Times: https://www.irishnews.com/news/northernirelandnews/2020/07/09/news/legal-proceedings-against-uk-government-over-awarding-of-108m-ppe-contracts-to-antrim-firm-1999417/

Ash, E., Galletta, S., & Giommoni, T. (2020). A machine learning approach to analyze and support anti-corruption policy. *Unpublished Manuscript*, 1-34.

Baranek, B. and Titl, V. (2020). 'The cost of favoritism in public procurement', *Journal of Law & Economics*.

Basu, S., Kumbier, K., Brown, J. B. and Yu, B. (2018). 'Iterative random forests to discover predictive and stable high-order interactions', *Proceedings of the National Academy of Sciences*, Vol. 115, pp. 1943–1948.

de Blasio, G., D'Ignazio, A., & Letta, M. (2020). *Predicting Corruption Crimes with machine learning*. *A study for the Italian Municipalities*. DiSSE Sapienza Working Paper Series No. 16/2020, 1–36.

Blumenstock, J. E. (2016). 'Fighting poverty with data: machine learning algorithms measure and target poverty', *Science*, Vol. 353, pp. 753–754.

Brayne, S. (2017). 'Big data surveillance: the case of policing', *American Sociological Review*, Vol. 82, pp. 977–1008.

Breiman, L. (2001). 'Random forests', *Machine Learning*, Vol. 45, pp. 5–32.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J. and Mullainathan, S. (2016). 'Productivity and Selection of Human Capital with Machine Learning', *American Economic Review*, Vol. 106, pp. 124–127.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, P. W. (2002). 'SMOTE: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357.

Cingano, F. and Pinotti, P. (2013). 'Politicians at work: the private returns and social costs of political connections', *Journal of the European Economic Association*, Vol. 11, pp. 433–465.

Claessens, S., Feijen, E. and Laeven, L. (2008). 'Political connections and preferential access to finance: the role of campaign contributions', *Journal of Financial Economics*, Vol. 88, pp. 554–580.

Colonnelli, E., Prem, M. and Teso, E. (2020). 'Patronage and selection in public sector organizations', *American Economic Review*, Vol. 110, pp. 3071–3099.

Conn, D., & Evans, R. (2020, December 3). *The Guardian*. Retrieved from Covid-19 contracts: government refuses to say who benefited from political connections: https://www.theguardian.com/world/2020/dec/03/government-secrecy-over-huge-covid-contracts-completely-unnecessary-say-critics

Decarolis, F. and Giorgiantonio, C. (2022). 'Corruption red flags in public procurement: new evidence from Italian calls for tenders', *EPJ Data Science*, Vol. 11, pp. 1–34.

Faccio, M. (2006). 'Politically connected firms', *American Economic Review*, Vol. 96, pp. 369–386.

Faccio, M. (2010). 'Differences between politically connected and nonconnected firms: a cross-country analysis', *Financial Management*, Vol. 39, pp. 905–927.

Ferwerda, J., Deleanu, I., & Unger, B. (2017). 'Corruption in public procurement: finding the right indicators', *European Journal on Criminal Policy and Research*, Vol. 23, pp. 245–267. https://doi.org/10.1007/s10610-016-9312-3

Fisman, R. and Wang, Y. (2015). 'The mortality cost of political connections', *Review of Economic Studies*, Vol. 82, pp. 1346–1382.
Friedman, J. H. (2001). 'Greedy function approximation: a gradient boosting machine', *Annals of Statistics*, Vol. 29, pp. 1189–1232.
Gallego, J., Rivero, G. and Martinez, J. (2021). 'Preventing rather than punishing: an early warning model of malfeasance in public procurement', *International Journal of Forecasting*, Vol. 37, pp. 360–377.
de Goeij, H., & Santora, M. (2019, June 23). *In the Largest Protests in Decades, Czechs Demand Resignation of Prime Minister*. Retrieved from *The New York Times*. https://www.nytimes.com/2019/06/23/world/europe/czech-republic-protests-andrej-babis.html
Harding, L. (2016, April 5). What are the Panama Papers? *A Guide to History's Biggest Data Leak*. Retrieved from The Guardian: https://www.theguardian.com/news/2016/apr/03/what-you-need-to-know-about-the-panama-papers
Ho, T. K. (1995). 'Random decision forests', in *Proceedings of 3rd International Conference on Document Analysis and Recognition* Vol. 1, pp. 278–282, Institute of Electrical and Electronics Engeneering, Montreal, Canada.
Hoerl, A. E. and Kennard, R. W. (1970). 'Ridge Regression: Biased Estimation for Nonorthogonal Problems', *Technometrics*, Vol. 12, pp. 55–67.
Hughes, A. (2017, May 17). *5 facts about U.S. political donations*. Retrieved from Pew Research Center: https://www.pewresearch.org/fact-tank/2017/05/17/5-facts-about-u-s-political-donations/
James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, Springer, New York.
Jean, N., Burke, M., Xie, M., Davis, M. W., Lobell, D. B. and Ermon, S. (2016). 'Combining satellite imagery and machine learning to predict poverty', *Science*, Vol. 353, pp. 790–794.
Khwaja, A. I. and Mian, A. (2005). 'Do lenders favor politically connected firms? rent provision in an emerging financial market', *Quarterly Journal of Economics*, Vol. 120, pp. 1371–1411.
Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z. (2015). 'Prediction policy problems', *American Economic Review: Papers & Proceedings*, Vol. 105, pp. 491–495.
Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. (2018). 'Human decisions and machine predictions', *Quarterly Journal of Economics*, Vol. 133, pp. 237–293.
Kondo, S., Daisuke, M., Kengo, S., Miki, S., & Teppei, U. (2019). *Using Machine Learning to Detect and Forecast Accounting Fraud*. RIETI Discussion Paper Series No. 19-E-103, pp. 1–61.
Lima, M. S. and Delen, D. (2020). 'Predicting and explaining corruption across countries: A machine learning approach', *Government Information Quarterly*, Vol. 37, pp. 1–15.
Lopez-Iturriaga, F. J. (2018). 'Predicting public corruption with neural networks: an analysis of Spanish provinces', *Social Indicators Research*, Vol. 140, pp. 975–998.
Mastrobuoni, G. (2020). 'Crime is terribly revealing: information technology and policy productivity', *Review of Economic Studies*, Vol. 87, pp. 2727–2753.
Meijer, A. and Wessels, M. (2019). 'Predictive policing: review of benefits and drawbacks', *International Journal of Public Administration*, Vol. 42, pp. 1031–1039.
Mohler, G. O., Short, M., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L. and Brantingham, P. J. (2015). 'Randomized controlled field trials of predictive policing', *Journal of the American Statistical Association*, Vol. 110, pp. 1399–1411.
Mullainathan, S. and Spiess, J. (2017). 'Machine learning: an applied econometric approach', *Journal of Economic Perspectives*, Vol. 31, pp. 87–106.
Murdoch, J. W., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B. (2019). 'Definitions, methods, and applications in interpretable machine learning', *Proceedings of the National Academy of Sciences of USA*, Vol. 116, pp. 22071–22080.
National Audit Office. (2020, November 26). *Investigation into Government Procurement during the COVID-19 Pandemic*. Retrieved from National Audit Office. https://www.nao.org.uk/report/government-procurement-during-the-covid-19-pandemic/
Observatory of Public Sector Innovation. (2016). *DOZORRO*. Retrieved from Observatory of Public Sector Innovation. https://oecd-opsi.org/innovations/dozorro/

Pegg, D., Lawrence, F., & Conn, D. (2020, November 18). *PPE Suppliers with Political Ties given 'High-Priority' Status, Report Reveals*. Retrieved from The Guardian. https://www.theguardian.com/politics/2020/nov/18/ppe-suppliers-with-political-ties-given-high-priority-status-report-reveals

Puterman, E., Weiss, J., Hives, B. A., Gemmill, A., Karasek, D., Mendes, W. B. and Rehkopf, D. H. (2020). 'Predicting mortality from 57 economic, behavioral, social, and psychological factors', *Proceedings of the National Academy of Sciences*, Vol. 117, pp. 16273–16282.

Rubin, R. (2020, January 6). *IRS Personal Income-Tax Audits Drop to Lowest Level in Decades*. Retrieved from The Wall Street Journal. https://www.wsj.com/articles/irs-personal-income-tax-audits-drop-to-lowest-level-in-decades-11578352541

Sukhtankar, S. (2012). 'Sweetening the deal? political connections and sugar mills in India', *American Economic Journal: Applied Economics*, Vol. 4, pp. 43–63.

Tibshirani, R. (1996). 'Regression shrinkage and selection via the Lasso', *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, pp. 267–288.

Titl, V. and Geys, B. (2019). 'Political donations and the allocation of public procurement contracts', *European Economic Review*, Vol. 111, pp. 443–458.

Titl, V., Geys, B. and De Witte, K. (2021). 'Political donations, public procurement and government efficiency', *World Development*, Vol. 148. https://doi.org/10.1016/j.worlddev.2021.105666

Transparency International. (2019). *Corruption Perceptions Index*. Retrieved from Transparency International. https://www.transparency.org/en/cpi

Wheeler, A. P. and Steenbeek, W. (2020). 'Mapping the risk terrain for crime using machine learning', *Journal of Quantitative Criminology*, Vol. 37, pp. 1–36.

Supporting Information

Additional Supporting Information may be found in the online Appendix:

Table S1. Predicting Political Connections Using Logistic Regression
Figure S1. Feature Importance Ranking for Boosting. *Note:* This graphs shows the feature importance for the most important variables from the boosting model used in the main estimation
Figure S2. Partial Dependence Plot. *Note:* This figure shows partial dependence plot showing the relationship between the two strongest predictors (i.e. firms' age and the value of public procurement contract) and the likelihood of being connected
Table S2. Confusion Matrices
Figure S3. Sensitivity of Predicting Political Donors Using Machine Learning. *Note:* The figure can be interpreted as follows: if 5,000 donating firms and 5,000 not donating firms are used to train the algorithm, boosting predicts 93.4% of donating firms correctly in a subsample of the same size randomly drawn from the rest of the sample
Figure S4. Specificity of Predicting Political Donors Using Machine Learning. *Note:* The figure can be interpreted as follows: if 5,000 connected firms and 5,000 not connected firms are used to train the algorithm, boosting predicts 94.5% of not-donating firms correctly in a subsample of the same size randomly drawn from the rest of the sample
Figure S5. Sensitivity of Predicting CEO Donors Using Machine Learning. *Note:* The figure can be interpreted as follows: if 5,000 firms with their CEOs donating and 5,000 firms with their not-donating CEOs are used to train the algorithm, boosting predicts 95.8% firms with donating CEOs correctly in a subsample of the same size randomly drawn from the rest of the sample

Figure S6. Specificity of Predicting CEO Donors Using Machine Learning. *Note:* The figure can be interpreted as follows: if 5,000 firms with their CEOs donating and 5,000 firms with their not-donating CEOs are used to train the algorithm, boosting predicts 98.3% of firm not-donating CEOs correctly in a subsample of the same size randomly drawn from the rest of the sample

Figure S7. Sensitivity of Predicting Personal Political Connections Using Machine Learning. *Note:* The figure can be interpreted as follows: if 5,000 personally connected firms and 5,000 not-connected firms are used to train the algorithm, boosting predicts 96% of personally politically connected firms correctly in a subsample of the same size randomly drawn from the rest of the sample

Figure S8. Specificity of Predicting Personal Political Connections Using Machine Learning. *Note:* The figure can be interpreted as follows: if 5,000 personally connected firms and 5,000 not-connected firms are used to train the algorithm, boosting predicts 84.2% of not-connected firms correctly in a subsample of the same size randomly drawn from the rest of the sample

Figure S9. Accuracy of Predicting Political Connections Using Machine Learning Using the Dataset from 2011. *Note:* The figure can be interpreted as follows: if 5,000 connected firms and 5,000 not connected firms are used to train the algorithm, boosting is 84.1% accurate in predicting which firms are politically connected on a subsample of the same size randomly drawn from the rest of the sample. The figure displays 95% confidence intervals. In this figure, the dataset of Czech firms from 2011 is used

Figure S10. Sensitivity of Predicting Political Connections Using Machine Learning Using the Dataset from 2011. *Note:* The figure can be interpreted as follows: if 5,000 connected firms and 5,000 not connected firms are used to train the algorithm, boosting predicts 87.2% of politically connected firms correctly in a subsample of the same size randomly drawn from the rest of the sample. In this figure, the dataset of Czech firms from 2011 is used

Figure S11. Specificity of Predicting Political Connections Using Machine Learning Using the Dataset from 2011. *Note:* The figure can be interpreted as follows: if 5,000 connected firms and 5,000 not connected firms are used to train the algorithm, boosting predicts 81% of not-politically connected firms correctly in a subsample of the same size randomly drawn from the rest of the sample. In this figure, the dataset of Czech firms from 2011 is used

Figure S12. Feature Importance Ranking for Boosting Using the Dataset from 2011. *Note:* This graphs shows the feature importance for the most important variables from the boosting model used in the main estimation

Figure S13. Politically Connected and Not Connected Firms According to Firm Profit Margin and Operating EBIT; Sample of Firms in Retail Sector

Figure S14. Politically Connected and Not Connected Firms According to Firm Profit Margin and Operating EBIT; Sample of Firms in Real Estate Sector

Figure S15. Politically Connected and Not Connected Firms According to Firm Profit Margin and Operating EBIT; Sample of Manufacturing Firms

Data replication package: the data replication package is available at https://doi.org/10.3886/E188961