Mazrekaj, Deni; Titl, Vitezslav; Schiltz, Fritz

Working Paper

Identifying Politically Connected Firms: A Machine Learning Approach

U.S.E. Working Papers Series, No. 21-10

Provided in Cooperation with:
Utrecht University School of Economics (U.S.E.), Utrecht University

Suggested Citation: Mazrekaj, Deni; Titl, Vitezslav; Schiltz, Fritz (2021) : Identifying Politically Connected Firms: A Machine Learning Approach, U.S.E. Working Papers Series, No. 21-10, Utrecht University, Utrecht University School of Economics (U.S.E.), Utrecht

This Version is available at:
https://hdl.handle.net/10419/323029

Utrecht University

School of Economics

Identifying Politically Connected Firms: A Machine Learning Approach

Deni Mazrekaj, Vitezslav Titl, Fritz Schiltz

U.S.E. Working Paper Series nr: 21-10

School of Economics

Utrecht University School of Economics (U.S.E.) is part of the faculty of Law, Economics and Governance at Utrecht University. The U.S.E. Research Institute focuses on high quality research in economics and business, with special attention to a multidisciplinary approach. In the working papers series the U.S.E. Research Institute publishes preliminary results of ongoing research for early dissemination, to enhance discussion with the academic community and with society at large.

The research findings reported in this paper are the result of the independent research of the author(s) and do not necessarily reflect the position of U.S.E. or Utrecht University in general.

U.S.E. Research Institute
Kriekenpitplein 21-22, 3584 EC Utrecht, The Netherlands Tel: +31 30 253 9800, e-mail: use.ri@uu.nl www.uu.nl/use/research

Utrecht University School of Economics

U.S.E. Research Institute
Working Paper Series 21-10
ISSN: 2666-8238

Identifying Politically Connected Firms: A Machine Learning Approach

Deni Mazrekaj[1]
Vitezslav Titl[2]
Fritz Schiltz[3]
[1] Department of Sociology, University of Oxford
[2]Utrecht University School of Economics, Utrecht University
[3] Leuven Economics of Education Research, KU Leuven

June 2021

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. We propose that machine learning algorithms should be used by public institutions to identify politically connected firms with potentially large conflicts of interests, and we provide easy to implement R code to replicate our results.

Keywords: Political Connections, Corruption, Prediction, Machine Learning

JEL classification: D72, D73, P16

Acknowledgements:
We would like to thank Benny Geys, Kristof De Witte, Giovanna D'Inverno, Mark Verhagen, and Aniek Sies for useful comments and suggestions. 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 financial support from the Czech Science Foundation – GACR through its project N. 21-15480M „The Role of Institutional Factors and Information in Public Procurement Markets". The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

Comments welcomed to: v.titl@uu.nl

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 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 health care 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 & 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, & 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, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018), Post Traumatic Stress Disorder (Abbasi, 2019), and even mortality (Puterman, et al., 2020). Ranking among the most corrupt countries in Europe according to the 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 the Czech Prime Minister Andrej Babis had significant conflicts of interests related to his private businesses. Specifically, his businesses received almost 20 million euros of EU agricultural subsidies while being Prime Minister (de Goeij & 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 United States have introduced a ban on corporate donations to political parties and information on firms' ownership structure and management are not available.[1] In Czechia, however, political donations are allowed, and firms' ownership structure and management can be retrieved. By employing 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

[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.

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.

Politically connected firms generate substantial economic and welfare costs for the society, reaching up to 1.9% of GDP every year (Khwaja & Mian, 2005). 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 & Pinotti, 2013; Colonnelli, Prem, & Teso, 2020; Fisman & Wang, 2015; Titl & Geys, 2019). Despite these negative implications of political connections, both firms and politicians have an incentive to become politically connected (Cingano & Pinotti, 2013; Faccio, 2006; Sukhtankar, 2012). 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, Ludwig, Mullainathan, & Obermeyer, 2015), and on 'predictive policing' in particular, the idea that criminal activities can be predicted and therefore prevented before they happen (Brayne, 2017; Meijer & 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 United States and the United Kingdom predicted 1.4 to 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 a 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) analyze cross-country data to predict and explain corruption across countries. Lopez-Ituriaga and Sanz (2018) use 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, Gelletta, and Giommoni (2020) 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, & Martinez, 2021) and in Italy (Decarolis & Giorgiantonio, 2020). We build on this literature by employing machine learning algorithms at the level of actual political connections.

Data

Our data include all firms registered in Czechia in 2018. 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.[2] Using company identifiers, we merged these two sources of political donations together by matching all donations made by firms to political parties as well as the exact amounts. 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.[3] We matched the lists of individual persons who donated with the lists of 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

[2] Accessible at https://www.udhpsh.cz/
[3] Accessible at https://portal.justice.cz/Justice2/Uvod/uvod.aspx

on (supervisory) board members that ran for political offices was created by matching elections' candidate lists[4] and the lists of board members of all Czech companies mentioned above. Part of the data on political donations and partially also personal connections are now available on a website PolitickeFinance.cz maintained by Datlab Institute.[5] The complete assembled political connections data will be made available upon request.

The data on predictors were obtained from Orbis database collected by Bureau van Dijk. This database provides standardised annual accounts (consolidated and unconsolidated), financial ratios, sectoral activities, and ownership data. We use all suitable variables included in the Orbis database as predictors of political connections.[6] According to the 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 much more complete than datasets from other countries covered by the database such as the United Kingdom or Germany. Moreover, merging these financial data with our self-compiled political connections data was straightforward by 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 ran by the Ministry of Regional Development.[7] 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

[4] Accessible at https://www.volby.cz/

[5] Accessible at http://www.politickefinance.cz/

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

[7] Accessible at https://www.mmr.cz/cs/uvod and http://www.isvz.cz/ISVZ/Podpora/ISVZ.aspx.

connections as an indicator given 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. 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**.

TABLE 1: DESCRIPTIVE STATISTICS

Predictor Variable	Mean	Std. Dev.	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	12.0	93.3	0	10,000
Number of director managers	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	1,953.1	328.9	0	2,018
Based in Prague	0.3	0.5	0	1
Sector (categorical)				
Politically connected	0.05	0.2	0	1

Methods

To predict political connections, we start from 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 will be provided upon request. 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, Witten, Hastie, & Tibshirani, 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, D'Ignazio, & Letta, 2020) and use the Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). This technique essentially randomly undersamples the majority class, i.e., 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. We perform SMOTE using the *DMwR* package in R.

Although randomly dividing the sample in 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 performed 10-Fold Cross Validation in R using the *boot* package.

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 the 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 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-a-versa from a policy perspective.

Logistic Regression

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

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

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. We perform logistic regression using the *glm* package in R.

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, 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 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 $\lambda \sum_{j=1}^{p} |\beta_j^2|/2$ 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, ridge regression is applied after standardizing the predictors. The standardization is done by default by the *SuperLearner* package.

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 advantage by maximizing with the log-likelihood with the following shrinkage penalty $\lambda \sum_{j=1}^{p} |\beta_j|$. With LASSO, coefficients 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. 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.

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 (Basu, Kumbier, Brown, & Yu, 2018; Mullainathan & Spiess, 2017). 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. Consider in **Figure 4** a fictitious example in which we classify 100 firms as being politically connected or not connected using only three variables: the number of employees, the operational result, and registered capital. A relatively large firm (more than 10,000 employees), with a strong operational result (more than 150 million euros), and less than 1 billion euros in registered capital has a higher probability to be not connected (8/15). Therefore, this firm will be classified as not connected.

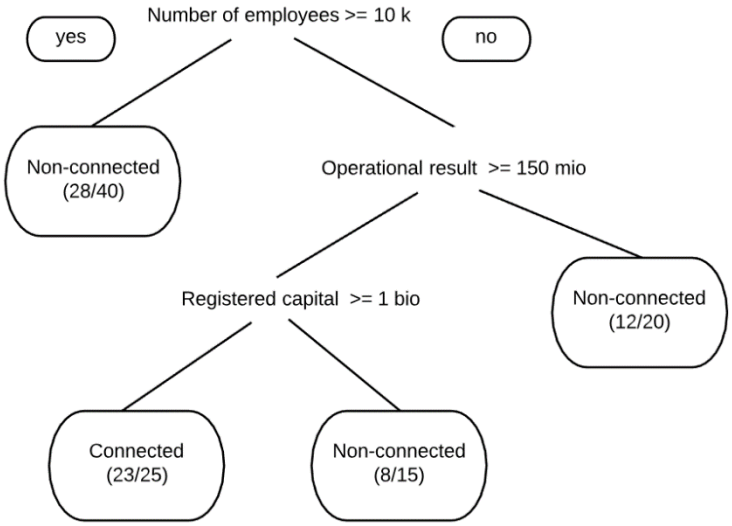

FIGURE 4: A FICTITIOUS EXAMPLE OF A CLASSIFICATION TREE FOR POLITICAL CONNECTIONS

Note: A relatively large firm (more than 10,000 employees), with a strong operational result (more than 150 million euros), and less than 1 billion euros in registered capital has a higher probability to be not connected (8/15) and will be classified as not connected.

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. 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.

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. 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.

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 1** 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.

Figure 1 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-a-versa. **Figure 2** shows the true positive rate, namely the rate at which politically connected firms are correctly predicted (*sensitivity*). **Figure 3** shows the true negative rate, the rate at which not politically connected firms are correctly predicted (*specificity*). 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 in predicting politically not connected firms, compared to the other methods. All five methods exhibit similar levels of accuracy with regard to predicting politically connected firms, although the variation in the accuracy levels appears higher.

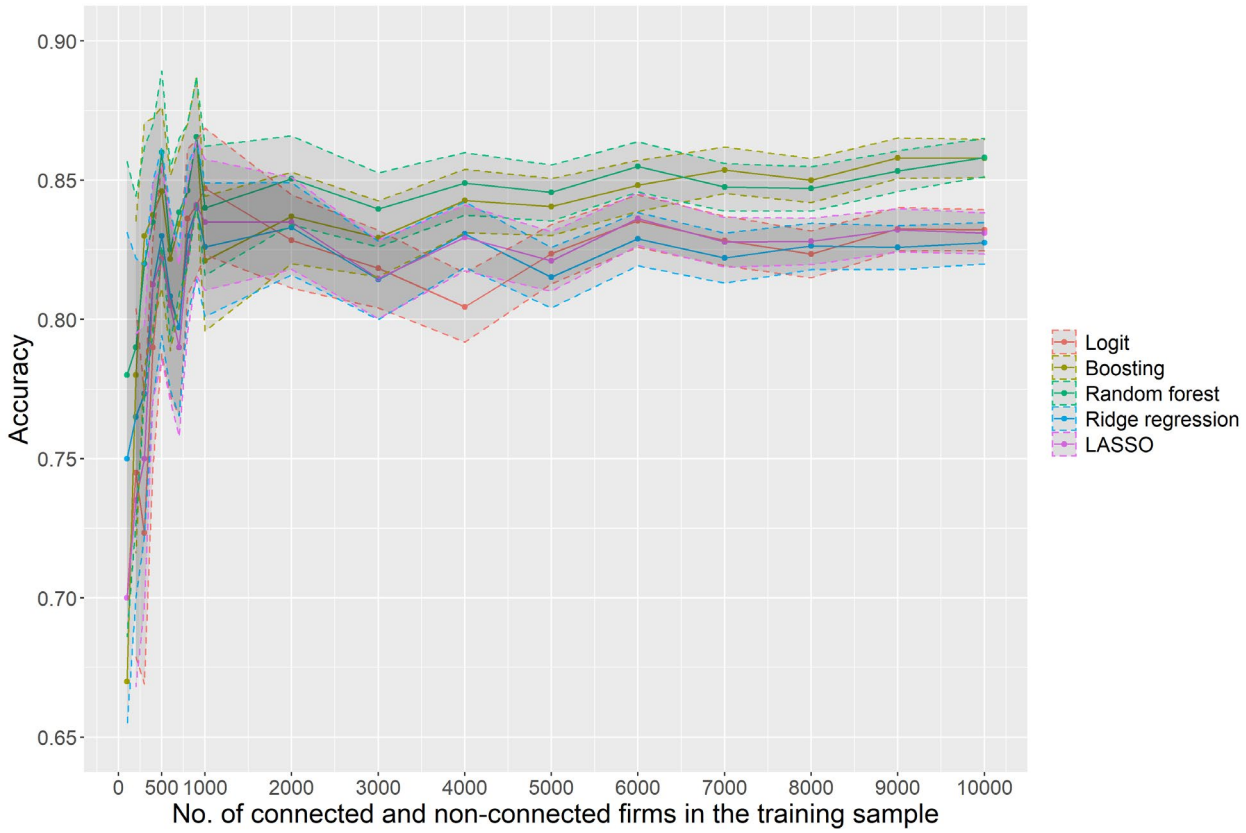

FIGURE 1: ACCURACY OF PREDICTING POLITICAL CONNECTIONS USING MACHINE LEARNING

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 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.

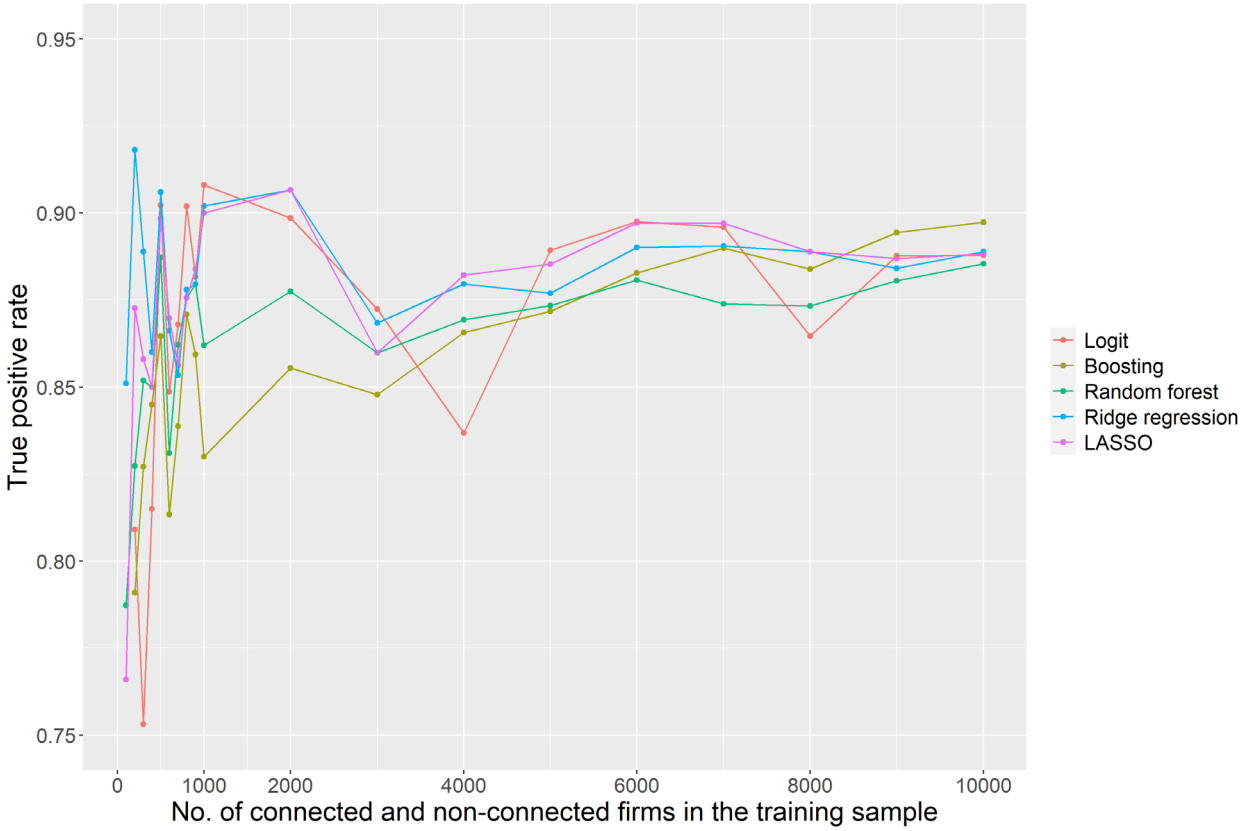

FIGURE 2: SENSITIVITY OF PREDICTING POLITICAL CONNECTIONS USING MACHINE LEARNING

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.

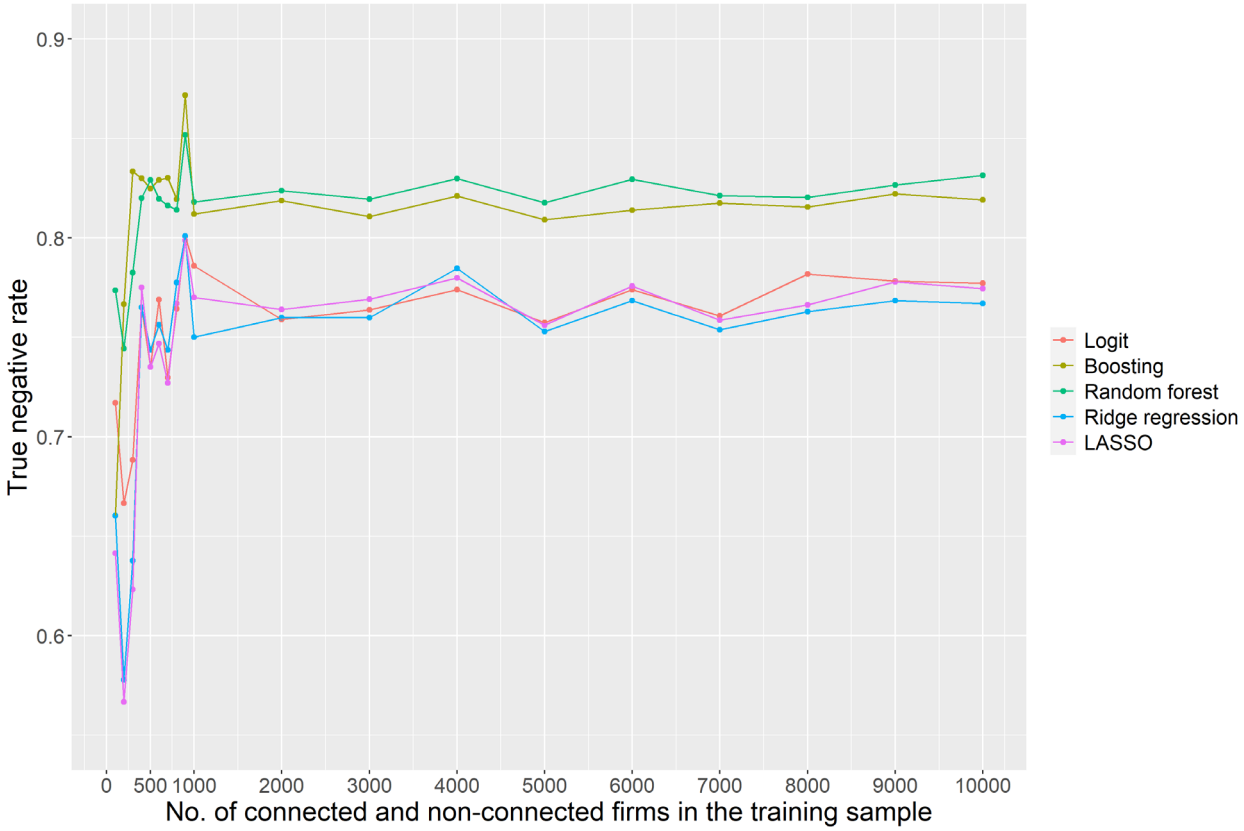

FIGURE 3: SPECIFICITY OF PREDICTING POLITICAL CONNECTIONS USING MACHINE LEARNING

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.

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 interests. Firms identified by the algorithms as politically connected can be targeted for inspection. Thereby, we could avoid welfare losses coming with politically connected firms (reaching up to 1.9% of GDP every year according to Khwaja & 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 & Pinotti, 2013; Colonnelli, Prem, & Teso, 2020; Fisman & Wang, 2015; Titl & Geys, 2019). 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 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. Over time, we expect firms to become familiar with machine learning algorithms and firms may improve their ability to fool the algorithm. Random audits would ensure that even the low-risk firms have some chance of being inspected. It is also useful to maintain some random auditing to persistently update the algorithm parameters for further targeted inspections (Ash, Galletta, & Giommoni, 2020). Further

studies should analyze 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.

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, Singh, Kumbier, Abbasi-Asl, & Yu, 2019), but in return achieve an 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. Nonetheless, our results indicate that even relatively easily interpretable algorithms such as random forests can predict political connections with a high accuracy.

References

Abbasi, J. (2019). First Biomarker-Based Screening Tool for PTSD. *Journal of the American Medical Association, 322*(15), 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.

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

Blumenstock, J. E. (2016). Fighting poverty with data: Machine learning algorithms measure and target poverty. *Science, 353*(6301), 753-754.

Brayne, S. (2017). Big Data Surveillance: The Case of Policing. *America Sociological Review, 82*(5), 977-1008.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and Selection of Human Capital with Machine Learning. *American Economic Review, 106*(5), 124-127.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, P. W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Cingano, F., & Pinotti, P. (2013). Politicians at work: the private returns and social costs of political connections. *Journal of the European Economic Association, 11*(2), 433-465.

Colonnelli, E., Prem, M., & Teso, E. (2020). Patronage and Selection in Public Sector Organizations. *American Economic Review, 110*(10), 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

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, 16/2020*, 1-36.

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

Decarolis, F., & Giorgiantonio, C. (2020). Corruption red flags in public procurement: new evidence from Italian calls for tenders. *Questioni di Economia e Finanza, Occasional Papers, 544*, 1-34.

Faccio, M. (2006). Politically Connected Firms. *American Economic Review, 96*(1), 369-386.

Fisman, R., & Wang, Y. (2015). The Mortality Cost of Political Connections. *Review of Economic Studies, 82*(4), 1346-1382.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics, 29*(5), 1189-1232.

Gallego, J., Rivero, G., & Martinez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting, 37*, 360-377.

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. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, pp. 278-282.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics, 12*(1), 55-67.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York: Springer.

Jean, N., Burke, M., Xie, M., Davis, M. W., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science, 353*(6301), 790-794.

Khwaja, A. I., & Mian, A. (2005). Do Lenders Favor Politically Connected Firms? Rent Provision in an Emerging Financial Market. *Quarterly Journal of Economics, 120*(4), 1371-1411.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *Quarterly Journal of Economics, 133*(1), 237-293.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review: Papers & Proceedings, 105*(5), 491-495.

Kondo, S., Daisuke, M., Kengo, S., Miki, S., & Teppei, U. (2019). Using Machine Learning to Detect and Forecast Accounting Fraud. *RIETI Discussion Paper Series, 19-E-103*, 1-61.

Lima, M. S., & Delen, D. (2020). Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly, 37*(101407), 1-15.

Lopez-Iturriaga, F. J. (2018). Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces. *Social Indicators Research, 140*, 975-998.

Mastrobuoni, G. (2020). Crime is Terribly Revealing: Information Technology and Policy Productivity. *Review of Economic Studies, 87*, 2727-2753.

Meijer, A., & Wessels, M. (2019). Predictive Policing: Review of Benefits and Drawbacks. *International Journal of Public Administration, 42*(12), 1031-1039.

Mohler, G. O., Short, M., Malinowski, S., Johnson, M., E., T. G., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, 1399-1411.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives, 31*(2), 87-106.

Murdoch, J. W., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *PNAS, 116*(44), 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., & Rehkopf, D. H. (2020). Predicting mortality from 57 economic, behavioral, social, and psychological factors. *Proceedings of the National Academy of Sciences, 117*(28), 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, 4*(3), 43-63.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267-288.

Titl, V., & Geys, B. (2019). Political Donations and the Allocation of Public Procurement Contracts. *European Economic Review, 111*, 443-458.

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

Wheeler, A. P., & Steenbeek, W. (2020). Mapping the Risk Terrain for Crime Using Machine Learning. *Journal of Quantitative Criminology*, 1-36.