Technological Forecasting & Social Change 184 (2022) 122016

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Gotham city. Predicting 'corrupted' municipalities with machine learning

Guido de Blasio [a], Alessio D'Ignazio [a], Marco Letta [b,*]

[a] Structural Economic Analysis Directorate, Bank of Italy, Italy
[b] Department of Social Sciences and Economics, Sapienza University of Rome, Italy

ARTICLE INFO

JEL classification:
C52
D73
H70
K10

Keywords:
Crime forecasting
White-collar crimes
Machine learning
Classification trees
Policy targeting

ABSTRACT

The economic costs of white-collar crimes, such as corruption, bribery, embezzlement, abuse of authority, and fraud, are substantial. How to eradicate them is a mounting task in many countries. Using police archives, we apply machine learning algorithms to predict corruption crimes in Italian municipalities. Drawing on input data from 2011, our classification trees correctly forecast over 70 % (about 80 %) of the municipalities that will experience corruption episodes (an increase in corruption crimes) over the period 2012–2014. We show that algorithmic predictions could strengthen the ability of the 2012 Italy's anti-corruption law to fight white-collar delinquencies and prevent the occurrence of such crimes while preserving transparency and accountability of the policymaker.

> "Corruption is widespread throughout Italy and represents one of the greatest obstacles to its growth, not only in civil terms but also in social and economic ones. Identifying the areas most exposed to corruption – with specific relation to different regional features – and drafting an Italian map of bribery is an essential tool to fight it."
> Raffaele Cantone, *President of the Italian Anti-corruption Authority* from 28 April 2014 to 23 October 2019.
> Trento, Italy, June 4, 2016

1. Introduction

Corruption in Italy is a major problem. According to the index of Transparency International,[1] which ranks 180 countries inversely by their perceived levels of public sector corruption as stated by experts and businesspeople, Italy was in the 51st position in 2019, far behind Germany (9th), France (23rd), and Spain (30th). Corruption used to be unevenly distributed across the country, but that is not the case anymore: traditionally linked with the presence of organized crime in the South, bribery has recently moved north both because organized crime converted from illegal activities to "normal" entrepreneurial business, and due to the emergency of clientelism and graft as cornerstones of the country's political establishment (see, for instance, Mocetti and Rizzica, 2019). As underscored by the former President of the Italian Anti-corruption Authority (quoted above), drafting a map of the areas where to concentrate investigation efforts should be considered a policy priority.

This paper provides such a map by means of machine learning (ML) algorithms. ML techniques have been developed in computer science and statistical literature and provide a powerful toolbox to deal with predictive tasks (Varian, 2014). In particular, their focus is on minimizing the out-of-sample prediction error and generalizing well on future unseen data (Athey and Imbens, 2017, 2019; Mullainathan and Spiess, 2017). We show that corruption is predictable, with high accuracy, using a limited set of variables easily available to policymakers. Moreover, we provide a simple rule for how to identify areas that will experience corruption episodes. Investigative and prosecution efforts should focus on these areas (rather than, as is currently the case, on presiding over areas that are not likely to have corruption problems).

We focus on white-collar crimes, which include, among others, corruption, fraud, and collusion. According to the FBI, the economic costs of such crimes are significantly larger than those associated with street crimes (Healy and Serafeim, 2016). Our prediction is based on the data taken from SDI ('Sistema d'Indagine'), the Ministry of Interior archive that contains records of all the crimes committed in the national

* Corresponding author.
E-mail addresses: guido.deblasio@bancaditalia.it (G. de Blasio), alessio.dignazio@bancaditalia.it (A. D'Ignazio), marco.letta@uniroma1.it (M. Letta).

[1] See https://www.transparency.org/en/cpi/2019/results/table.

https://doi.org/10.1016/j.techfore.2022.122016
Received 8 September 2021; Received in revised form 26 August 2022; Accepted 1 September 2022

territory at the municipality level. This dataset, derived from the IT system used by the police for investigation activities, has two major advantages: first, because it reports all the open cases which are under investigation by the police, it provides an instantaneous picture of the criminal activity in the municipality, whereas most datasets on crimes only report arrests or convictions which occur with a long delay with respect to when the crime is committed. Therefore, we are able to base our forecasts on an updated picture of the corruption activity going on over the country. If, instead, we had to use the data on arrests or convictions, our out-of-sample accuracy would have been worse given the delay. Second, our dataset is less subject to problems of underreporting of crimes because, on top of the reports filed by those affected by the crimes, it also contains records of all the investigations opened by the police forces themselves. This is a particularly valuable aspect in the case of corruption crimes: in such crimes, neither of the parties involved has any interest in reporting the crime because they would both be guilty of a criminal offence. The classification of crimes available in the SDI is made directly by the Ministry of Interior on the basis of the respective applicable law. We thus identify as white-collar crimes all crimes committed against articles 314–323 (crimes against public administration, such as embezzlement, grafting, and so on) and 479–481 (crimes against public faith, such as material or ideological fraud) of the Italian penal code: these include corruption, bribery, embezzlement, abuse of authority, and fraud.

Armed with the SDI data (and a large set of municipality-level features), we train and test our algorithms on the data referring to the period 2011–2012. Then, we evaluate the accuracy of the predictions by using data from 2012 to 2014. The results we present are based on a classification tree (Hastie et al., 2009). It is well known that the prediction accuracy of this algorithm might be inferior to that of other possible alternatives. However, we decide to focus on a classification tree because it provides the highest transparency concerning the variables chosen for the prediction. As we want to illustrate to public authorities the potential of algorithms to predict corruption, transparency is a fundamental requirement. Our results show that a classification tree provides a quite high out-of-sample prediction accuracy. Depending on the outcome variable, which can be specified in levels or variations, we are able to forecast from more than 70 % to almost 80 % of the local corruption. The prediction depends on the values of a few features, primarily referring to the characteristics of the local labour and housing markets, and the previous history of white-collar crimes. Crucially, we employ these findings to discuss how ML forecasts can be used to strengthen the effectiveness of anticorruption policy targeting. A benchmark to assess the potential gains in the fight against corruption stemming from the use of ML algorithms in Italy is provided by Law 190/2012. The law envisages a stricter anti-corruption regulation for all the municipalities with more than 15 thousand inhabitants. We show that such a threshold identifies only a fraction of the municipalities that have experienced corruption during the period 2012–2014, while ML predictions would have significantly improved anti-corruption efforts on the ground. In carrying out this comparison, we also discuss a number of issues related to the adoption of ML algorithms to fight corruption and improve law and policy targeting. For instance, we argue that our estimates at this stage are likely to provide only a very conservative approximation of the overall prediction gains attainable from ML.

Our contribution is particularly important for the policy makers. We show how some statistical tools, which can now be easily implemented by police statistical offices or by consultants hired for this purpose, using information that is in the public domain, in addition to that already held by the police, can predict the occurrence of white-collar crimes. We discuss how the advantages of using ML methods seem decisive, and we also illustrate potential drawbacks, such as those related to the transparency and cost-effectiveness of the proposed algorithmic predictions.

The paper is structured as follows. Section 2 provides a short overview of the literature. Section 3 highlights the data we use and provides a brief description of the ML methodology. The results are illustrated in Section 4. The comparison with the 2012 law is in Section 5. Section 6 concludes by offering a number of issues for discussion.

2. Literature review

Our research is related to two streams of literature: (i) the research focusing on the so-called policy prediction problems, i.e. applications where improved predictions can generate a large social impact of the policy; (ii) the research on crime, and in particular to that instrumental to predictive policing.

The first stream of research extensively relies on ML algorithms, and applications are already numerous in many fields. For instance, they include: predicting the riskiest patients for which a joint replacement would be futile (Kleinberg et al., 2015); improving judges' decisions on whether to detain or release arrestees as they await adjudication of their case (Kleinberg et al., 2018); targeting restaurant hygiene inspections (Kang et al., 2013); predicting highest risk youth for anti-violence interventions (Chandler et al., 2011); predicting the effectiveness of teachers in terms of value added (Rockoff et al., 2011); hiring police officers who will not behave violently, as well as promoting the best teachers only (Chalfin et al., 2016). ML techniques are also gradually becoming mainstream tools to improve poverty (Blumenstock et al., 2015; Jean et al., 2016; McBride and Nichols, 2018; Perez et al., 2019), resilience (Garbero and Letta, 2022), and food insecurity (Hossain et al., 2019; Knippenberg et al., 2019; Lentz et al., 2019) targeting; enhancing the effectiveness of public programs (Andini et al., 2018; Andini et al., 2022; Ballestar et al., 2019) and the understanding of food system dynamics (Garbero et al., 2021). As for the Italian context, recent works have leveraged the potential of ML to predict the bankruptcy of local governments (Antulov-Fantulin et al., 2021) and vaccine hesitancy in Italian municipalities (Carrieri et al., 2021).

The literature on crime known under the heading of ‘predictive policing’ builds on the idea that forecasting (and preventing) crime before it happens is necessary in order to reduce criminality and use public resources more efficiently (Brayne, 2017). Indeed, despite a great deal of randomness associated with occurrences of criminal episodes, there are patterns that can be detected (Gorr and Harries, 2003). The idea of being able to predict (and hence prevent) crime before it happens has gained increasing interest in the last few years across both researchers and police forces (Brayne and Christin, 2020; Meijer and Wessels, 2019). Among the statistical methods employed to forecast crime, two main approaches can be noted: a first set of methods emphasizes the spatial clustering of criminal activity and leads to the identification of the so-called hotspots, i.e., areas where offenders tend to repeat their crime (see Mohler, 2014); a second approach, boosted by the availability of big data and ML techniques, involves the identification of police targets through statistical predictions (Perry et al., 2013).

The latter stream of research is the one more closely related to our paper. It mainly focused on predictive policing based on crimes such as burglaries, thefts, and violence against the person (see Meijer and Wessels, 2019, and Bennett Moses and Chan, 2018, for a review). On the other hand, white-collar offences have been scantily studied. Among the papers focusing on white-collar crimes, López-Iturriaga and Sanz (2018) refer to the case of Spanish provinces and use info on corruption episodes to devise a neural network prediction model for corruption. Clifton et al. (2017) focus on another typical case of white-collar criminality, i.e., financial fraud, which is predicted by employing random forest algorithms on US local-level data. Ash et al. (2020) apply machine learning techniques, specifically tree-based gradient boosting, to detect Brazil's local-government corruption using budget accounts data. Lima and Delen (2020) employ a variety of ML algorithms, including random forests, support vector machines and artificial neural networks, on cross-country data to identify the most important predictors of corruption at the country level. Gallego et al. (2021) use a large micro dataset with more than 2 million public contracts to investigate the potential of ML to track and prevent corruption episodes in

public procurement in Colombia and understand its main drivers. Finally, Decarolis and Giorgiantonio (2020) use three machine learning routines, namely LASSO, ridge regression and random forest, on data concerning the procurement of public works to predict indicators of corruption risk, showing the potential of such algorithms in detecting corruption in public procurement.

We contribute to the literature on white-collar crimes by exploiting a rich dataset at the municipality-level in Italy. In particular, we can comprehensively delve into all kinds of white-collar crime episodes, including, as mentioned, embezzlement, grafting, corruption, bribery, abuse of authority, fraud, and others, and across all sectors.[2]

Machine learning tools are now widely implemented in predictive policing across the US, while in Europe, the use of algorithms in policing is at an embryonal level and mostly involves the United Kingdom.[3] Concerning Italy, Mastrobuoni (2020) studies predictive policing software used by the police department of Milan to study individual crime incidents, providing evidence of the substantial increase in police productivity guaranteed by the software. While most applications of predictive policing by means of ML involve violent crimes, ML tools aiming at preventing white-collar crimes are little used. A notable example refers to an application developed in Mexico in order to tackle corruption in public procurement and detect frauds from taxpayers: a pilot scheme showed that thanks to such algorithms it was possible to individuate fraudulent operations much more quickly (in about 1/3 of the time) than previous investigation methods.[4]

3. Data and methods

We first describe the data used for the ML predictive exercise (3.1) and then provide a short overview of the specific algorithms we employ (3.2).

3.1. Data

The source of crime data is the SDI archive by the Ministry of the Interior. Our database includes crime data for almost all Italian municipalities (8049 out of 8092 municipalities); data on white-collar crimes, which is the main object of our study, are available for 7794 municipalities over the years 2008–2014. In particular, we know the number of white-collar crimes for municipality and year, while the economic value of the crime and how many people are involved are unknown. The creation of the corruption outcome variables proceeds as follows. First, for each municipality in our sample, we divide the number of white-collar crimes in each given year by the corresponding total population in the same year. This way, we obtain a municipality-level corruption crime rate, in which the number of white-collar crimes, which tends to be mechanically higher in large municipalities, gets normalized by population levels. This indicator of white-collar crime rates is a continuous variable, which will take value zero in all the municipalities in which no white-collar crimes were reported in a given year. We then transform the continuous corruption crime rates into two separate corruption indicators capturing, respectively, whether a given municipality had corruption episodes in a given year, and whether a given municipality saw increases in corruption episodes with respect to the previous year. As we aim to compare the accuracy ML predictions with current regulations and targeting rules, we tackle corruption prediction as a classification rather than a regression problem.

More specifically, we predict two variables at the municipality level: (i) *WC crime rate*, a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) is positive, and 0 otherwise; (ii) *Δ WC crime rate*, a binary variable taking value 1 if the white-collar crime rate has increased with respect to the previous year, and 0 otherwise.[5] We compute these binary outcome variables for all years for which we have corruption data in our sample, namely between 2008 and 2014. Figs. 1 and 2 provide national maps for the value of our two target variables in the year 2012. The figures confirm that, nowadays, corruption is not a phenomenon that is exclusively predominant in the South of Italy: both the presence of corruption episodes and the increase in corruption levels over time are characterized by a scattered distribution across the entire Italian territory, with Apulia, Emilia, Lazio, and Sicily seemingly more affected than other regions.

The set of predictors consists of socio-economic, demographic, geographic and biophysical characteristics, all available at the municipality level. More specifically, we employ three different data sources to build the set of predictors: the 2011 version of the "8 Mila Census" dataset by the Italian National Institute of Statistics (ISTAT), which includes a wide variety of variables and indicators capturing socio-economic, labour market, demographic and housing market characteristics for Italian municipalities[6]; data on the number of foreign people by nationality, again drawn from ISTAT, from which we select the share of foreign people from the three most important origin regions (namely Southern Europe, Eastern Europe and Northern Africa); data on the number of police stations within 50 km from the municipality centroid and the Euclidean distance from the centroid to the closest police station[7]; climate data from the University of Delaware weather database (Matsuura and Willmott, 2015),[8] to control for biophysical and climatic heterogeneity across the country.[9] We also employ as additional predictors lagged (2008–2011) values of the outcome variables. As a result, we are able to assemble a dataset with almost 100 features. Table 1 shows summary statistics for a selected list of features, while Table A.1 in Appendix A reports the full list of predictors. The average municipality population is roughly equal to 7500 inhabitants, while the share of immigrants is lower than 6 %. The share of young people with higher education (19 %) is quite reduced compared to countries with a similar level of socio-economic development. There are significant gender and youth dimensions in the labour market. Police stations are quite widespread across the country, and the climate pattern (13 °C on average) is quite enjoyable. All of these variables refer to 2011. Our predictions for the years 2012–2014 will thus be static, as the values of the municipality-level indicators used as predictors will be fixed. Therefore, when we employ our machine learning algorithm to forecast corruption crimes, we will be using only information available to the Italian policymaker at the time of the ratification of the law, i.e., in 2012,

[2] Apart from their procurement focus, Gallego et al. (2021) and Decarolis and Giorgiantonio (2020) are micro-level analyses not aimed at forecasting the spatial distribution of white-collar crime occurrences. In contrast, our explicit aim is to provide an ML model able to map corruption episodes and thus help police investigations on the ground.

[3] In particular, Durham Constabulary has employed a risk assessment tool, constructed using random forests, to predict the risk of reoffending and used to decide whether some individuals should be prosecuted or not (Oswald et al., 2018).

[4] *Toward an AI strategy in Mexico. Harnessing the AI revolution.* Available at: https://7da2ca8d-b80d-4593-a0ab-5272e2b9c6c5.filesusr.com/ugd/7be025_e726c582191c49d2b8b6517a590151f6.pdf.

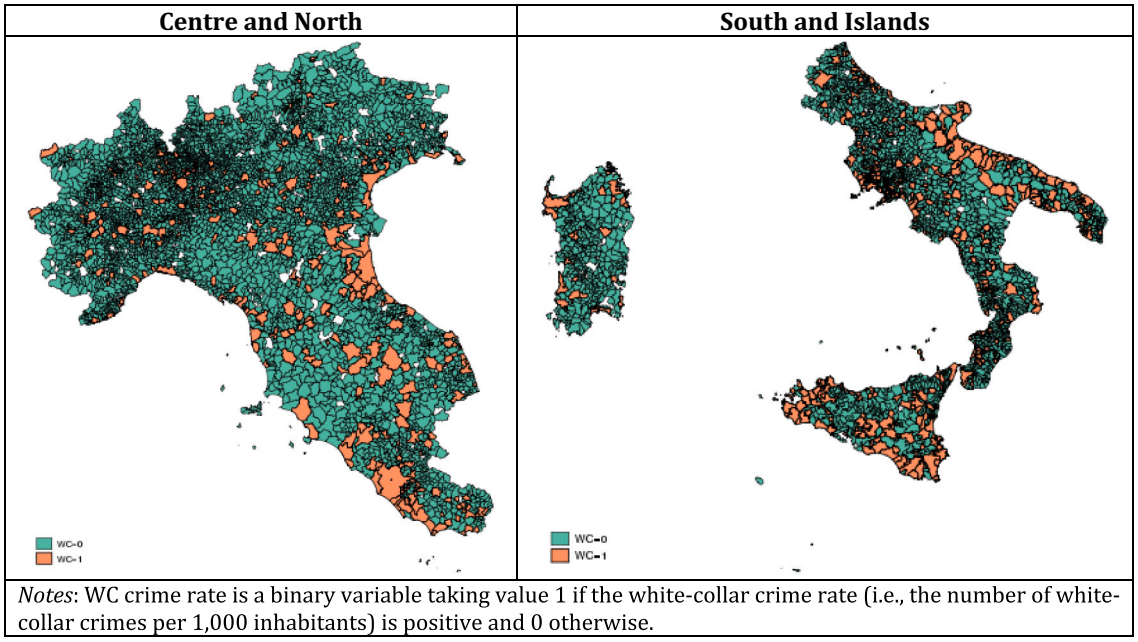
[5] As the data for the bulk of our features are not available before 2011, and since we want to put ourselves in the shoes of the policymaker in the year before the anti-corruption law, we use only 2012–2014 data on white-collar crimes to build our outcome variables and employ the data of the previous years as additional predictors capturing lagged crime rates.

[6] The 8 Mila Census database is publicly available at the following link: http://ottomilacensus.istat.it/.

[7] Freely available here.

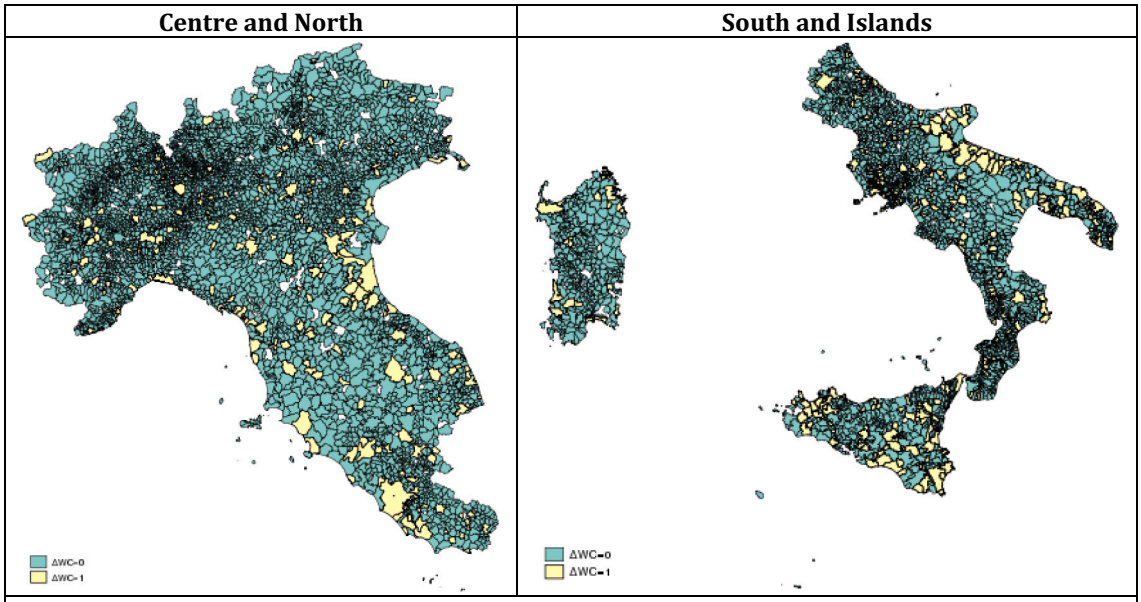
[8] The raw data on police stations at the local level are available here.

[9] A recent flourishing literature provides empirical evidence on the causal links between local weather and violent and non-violent crime trends. See, among the others, Horrocks and Menclova (2011), Ranson (2014), Chen et al. (2015) and Baysan et al. (2019).



Notes: WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1,000 inhabitants) is positive and 0 otherwise.

Fig. 1. WC crime rate across Italian municipalities – 2012
Notes: WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) is positive and 0 otherwise.



Notes: Δ WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1,000 inhabitants) has increased with respect to the previous year, and 0 otherwise.

Fig. 2. ΔWC crime rate across Italian municipalities – 2012
Notes: Δ WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) has increased with respect to the previous year, and 0 otherwise.

consistently with our aim of showing the ability of ML in informing the actual anti-corruption strategy, Table 2 provides descriptive statistics for the white-collar crime variables over the 2008–2014 period. It is interesting to note a sharp increase in white-collar crimes in 2011 compared to previous years: the number of municipalities with corruption episodes more than doubled. Despite slight decreases in the following years, corruption crimes are sensibly higher for the 2011–2014 period compared to 2008–2010. In sum, the available data confirm that the corruption problem in Italy is getting worse over time, and so are its social and economic costs.

4. Methods

ML techniques use highly flexible functional forms. The degree of flexibility is the result of a well-known bias-variance trade-off: allowing for more flexibility improves the in-sample fit at the cost of reducing the out-of-sample fit (over-fitting). In order to choose the optimal level of complexity, ML algorithms typically rely on empirical tuning. Following

Table 1
Descriptive statistics – selected features.

Variable	Year	Mean	Var	sd	Count
Population	2011	7472.740	1.637e+09	40,457.589	7794
Number of foreign people per 1000 inhabitants	2011	58.496	1776.420	42.148	7794
Average household size	2011	2.359	0.070	0.265	7794
Share of real estate ownership among households	2011	76.810	44.517	6.672	7794
Mean surface of inhabited buildings (sq. m.)	2011	103.235	171.943	13.113	7794
Share of buildings into disuse	2011	1.639	3.960	1.990	7794
Share of young people with a university degree	2011	18.873	56.436	7.512	7794
Share of adult people with secondary education	2011	38.219	42.002	6.481	7794
Male unemployment rate	2011	8.357	32.100	5.666	7794
Female unemployment rate	2011	12.940	65.844	8.114	7794
Unemployment rate	2011	10.184	40.106	6.333	7794
Youth unemployment rate	2011	29.357	232.692	15.254	7782
Daily mobility outside the municipality for study or work (share of the working-age population)	2011	35.067	160.360	12.663	7794
Daily student mobility outside the municipality (share of the population that moves daily outside the municipality)	2011	113.228	19,856.800	140.914	7071
Vulnerability index	2011	98.770	2.701	1.644	7794
Place in vulnerability index ranking	2011	4036.333	5,454,224.210	2335.428	7794
Share of households in potential economic hardships	2011	2.041	3.561	1.887	7794
Number of police stations within 50 km from the municipality centroid	2011	10.414	99.024	9.951	7794
Share of foreign people from Eastern Europe	2011	0.424	0.054	0.233	7794
Share of foreign people from Northern Africa	2011	0.158	0.025	0.157	7794
Share of foreign people from Southern Europe	2011	0.144	0.022	0.148	7794
Average temperature (°C)	2011	13.252	10.025	3.166	7794
Total precipitation (mm)	2011	859.429	96,898.239	311.285	7794

Table 2
Descriptive statistics – white-collar crime variables.

Variable	Year	Mean	Var	sd	Obs
WC crime rate	2008	0.0457	0.0436	0.209	7794
	2009	0.0499	0.0474	0.218	7794
	2010	0.0533	0.0504	0.225	7794
	2011	0.113	0.100	0.317	7794
	2012	0.0966	0.0873	0.295	7794
	2013	0.0920	0.0835	0.289	7794
	2014	0.0958	0.0867	0.294	7794
Δ WC crime rate	2008	0.0368	0.0355	0.188	7794
	2009	0.0331	0.0320	0.179	7794
	2010	0.0379	0.0364	0.191	7794
	2011	0.0908	0.0826	0.287	7794
	2012	0.0710	0.0659	0.257	7794
	2013	0.0429	0.0410	0.203	7794
	2014	0.0667	0.0623	0.250	7794

Notes: WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) is positive and 0 otherwise. Δ WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) has increased with respect to the previous year, and 0 otherwise.

a standard ML approach, we randomly split our sample into two sets, containing, respectively, 2/3 and 1/3 of municipalities. We use the first set to train our algorithms (training set), while we use the second to test them (testing set). The out-of-sample performance of the model on the unseen (held-out) data of the testing set can then be considered a reliable measure of the 'true' performance on future data. In order to solve the bias-variance trade-off (Hastie et al., 2009), we employ 10-fold cross-validation on the training sample to select the best-performing values of the key tuning parameter. While we train our model using pre-2012 features and 2012 outcomes, the predictive performance of our preferred algorithm is also evaluated over the years 2013 and 2014. To this aim, we decided to consider municipalities belonging to the testing set only. Employing our algorithm to predict 2013 and 2014 outcomes also for the municipalities in the training set would lead to an upward bias in the accuracy. The reason is that most of the features that would be used to predict 2013 and 2014 outcomes in the training set of municipalities refer to the year 2011, and hence they were already used to "learn" the model in the same set of municipalities.

Our preferred ML algorithm is the classification tree (Hastie et al., 2009). Classification trees provide the researcher with a clear scheme (the tree) to follow for targeting and are particularly suited for applications in which the decision rule needs to be transparent (Lantz, 2019), such as when the output of the model must be shared in order to facilitate public decision making (Andini et al., 2018). As it will be clear in Section 4, the output of a decision tree algorithm is intuitive and can be easily understood also by people without a strong statistical background, making it appealing for policy targeting purposes. From a technical point of view, the algorithm divides the data into progressively smaller subsets to identify patterns useful for predicting a specific

discrete output. Trees are highly flexible methods because non-linearities and interactions are easily captured by the sequence of splits. In principle, classification trees allow one to reach a perfect in-sample fit by adding more and more leaves, but, in practice, regularization via tree pruning and cross-validation is used to tune the best-performing hyperparameter. In fact, a high number of levels in a tree is likely to overfit the data, leading to a poor predictive model. The solution to this issue is to reduce the complexity of the tree by setting a complexity parameter (*cp*) and using it to prune the tree. We select the optimal value of *cp* via 10-fold cross-validation in the training dataset.

Before performing our classification exercise, we need to tackle the challenge stemming from our highly imbalanced dataset. In our sample corruption is, from a statistical point of view, a rare event: our outcome variables are both highly skewed toward zeros (see Figs. 1 and 2 and Table 2). In the case of imbalanced datasets, predictive algorithms run into the so-called "accuracy paradox": they provide predictions featured by a high out-of-sample accuracy (even greater than 90 %), but useless for practical purposes, because they simply predict the over-represented class (y = 0 in our case). This happens because the algorithms aim to provide the lowest *total* error rate, irrespective of which class the errors come from. Therefore, predictive exercises on imbalanced datasets tend to result in a very high specificity (i.e., the proportion of negatives cases correctly identified) but an extremely low, if not null, sensitivity (the proportion of positives cases correctly identified).

This is visible in our case by looking at the prediction accuracy we obtain by using the original sample: Tables A.2 and A.3 in Appendix A show that the imbalanced data cannot be employed to identify 'corrupted' municipalities successfully. With reference to 2012, the percentage of correctly predicted cases is greater than 90 %, both for *WC crime rate* and *Δ WC crime rate*, but the accuracy for the y = 1 cases is slightly above 30 % for the former outcome and 0 for the latter. To tackle this issue, we make use of the Synthetic Minority Oversampling Technique (SMOTE) routine developed by Chawla et al. (2002) to rebalance the two classes in our training sample. We implement the SMOTE algorithm only on the training subsample, leaving the testing sample untouched. This means that the training dataset is artificially balanced over the two outcomes, while the out-of-sample performance is evaluated on the original (skewed) testing set. After rebalancing our training dataset, the two outcomes are almost perfectly balanced between the two classes, and the sample size is smaller.[10] On these rebalanced data, we then apply our decision tree algorithm, whose results are provided in the next section.

In the methodological annex reported in Appendix B, we provide essential background information about the machine learning methodology and analytical tools employed in this study.

5. Results

Fig. 3 pictures the classification tree that predicts the probability that a given municipality experiences corruption crimes (i.e., the 2012 crime rate is greater than 0). The algorithm uses three predictors: municipality population, 2011 white-collar crime rate, and the share of the working-age population involved in daily extra-municipal mobility for study or work reasons (from now on, mobility share). For instance, municipalities with a population larger than 7390 residents are predicted as prone to corruption if their 2011 white-collar crime rate was higher than a given cutoff (0.0000349). If the 2011 crime rate was lower than that threshold, the algorithm takes into account the value of the mobility share and predict as potentially 'corrupted' the places with a mobility share lower than 39.6 %. Fig. 4 illustrates the classification tree for the outcome defined in variations. In this case, the decision tree selects predictors that refer to characteristics of the local labour and housing markets. For instance, the algorithm predicts an increase of white-collar crimes in municipalities with more than 7361 inhabitants, with a mobility share higher than 38 %, where buildings have less than 106 square meters on average, and the share of buildings in disuse is larger than 1.2 %. It is worth noticing that the fact that some particular socioeconomic characteristics are useful for prediction purposes does not imply that they represent determinants of corruptive practices (see Mullainathan and Spiess, 2017). In the case at hand, the fact that the algorithm uses city size and mobility size suggests that in smaller, closed municipalities, broader social control might work to prevent corruption.

Tables 3 and 4 highlight the prediction accuracy of the classification trees described in Figs. 3 and 4, respectively. We evaluate such performance for the years 2012, 2013 and 2014 for the municipalities belonging to the testing set only, as argued in Section 4. Remember that these predictions are static, as the model was trained using only predictors referring to the years before the ratification of the anti-corruption regulation (2012). Concerning the *WC crime rate*, overall accuracy is very high and consistently around 85 % for the three years (Table 3). Specificity is even higher. Sensitivity is lower, ranging from 72.2 % in 2014 to 74.3 % in 2012, but still quite high if compared with the pre-SMOTE performance of the algorithm, reported in Table A.2 (where sensitivity for the 2012 sample is 31.2 %, while the overall accuracy is 92.6 %).[11] The prediction accuracy for the tree described in Fig. 4, where the outcome is defined in variations, is slightly lower. The percentage of correctly predicted cases is always around 75 %, and the prediction accuracy related to the y = 1 cases is substantial (from 74 % to 80 %), and in some cases, even higher than specificity. For comparison, notice that the classification tree without SMOTE (Table A.3, which refers to 2012) would have delivered an overall accuracy of over 93 %, due to a 100 % accuracy for y = 0 and a 0 % accuracy for y = 1.[12]

ML algorithms use highly non-linear functional forms. However, endowed with the same set of predictors used to produce the classifications trees of Figs. 3 and 4, one can also run simpler logit regressions to gauge the magnitude of the accuracy gains due to more complex functional forms. Tables A.4 and A.5 provide such regression results for the two outcomes, respectively. We find that logit predictions drastically reduce sensitivity when the outcome is defined in variations, while when the outcome is *WC crime rate,* the benefits of more complexity involve the 2012 and 2014 predictions. On top of this, while classification trees automatically handle missing data (through the use of surrogate splits), logit does not, so its implementation required an ex-ante imputation of missing values.[13] Finally, the classification tree for *WC crime rate* uses the lagged (2011) WC crime rate to derive its predictions. This variable is taken from the SDI archive, and it is routinely available only to the police department. An interesting robustness test refers to the scenario where such variable is unavailable. We made the algorithm blind to the lagged (2008–2011) values of the crime variables taken from SDI. Results are depicted in Table A.6. We find only a limited reduction in overall accuracy. Unexpectedly, we observe that, when we

[10] The sample size is 1468 observations for the rebalanced training dataset using the *Δ WC crime rate* target variable, of which 722 negatives (y = 0) and 746 positives (y = 1); and 2033 observations for the rebalanced training dataset using the *WC crime rate* target variable, of which 993 negatives (y = 0) and 1040 positives (y = 1).

[11] The corresponding tree for the pre-SMOTE dataset is reported in Figure A.1. There are only two predictors in the tree: population and the extra-municipality mobility share.

[12] There is no corresponding tree for this table on the pre-SMOTE classification performance because there is no tree (the algorithm predicts all zeroes).

[13] We also implemented a simpler logit model with only a limited set of predictors typically associated with corruption crimes. Specifically, we ran a model with a similar vector of covariates (population, employment and unemployment rates, educational attainment variables) to that included in the specification adopted by De Angelis et al. (2020). In this case, the out-of-sample sensitivity performance of this basic model was substantially worse than that of the classification tree for both outcome variables in all years.

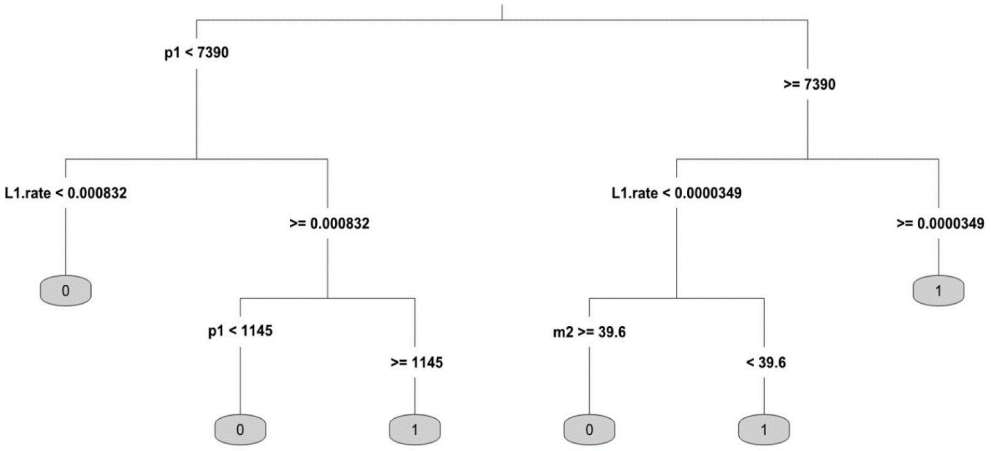


Fig. 3. Classification tree for WC Crime Rate – Post-SMOTE data
L1.rate: Lagged (2011) WC crime rate
p1: Population
m2: Daily mobility outside the municipality for study or work
(share of the working-age population).

Legend	
L1.rate:	Lagged (2011) WC crime rate
p1:	Population
m2	Daily mobility outside the municipality for study or work (share of the working-age population)

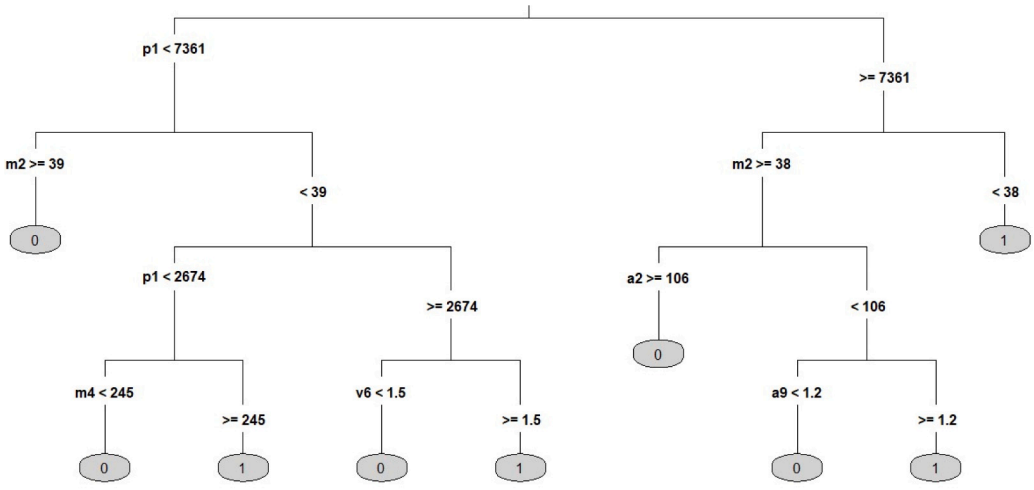


Fig. 4. Classification tree for Δ WC Crime Rate – Post-SMOTE data
p1: Population
m2: Daily mobility outside the municipality for study or work
(share of the working-age population)
a2: Mean surface of inhabited buildings (square meters)
m4: Daily student mobility outside the municipality
(share of the population that moves daily outside the municipality)
v6: Share of households in potential economic hardships (%)
a9: Share of buildings into disuse (%).

Legend	
p1:	Population
m2:	Daily mobility outside the municipality for study or work (share of the working-age population)
a2:	Mean surface of inhabited buildings (square meters)
m4:	Daily student mobility outside the municipality (share of the population that moves daily outside the municipality)
v6:	Share of households in potential economic hardships (%)
a9:	Share of buildings into disuse (%)

do not consider previous corruption episodes, the sensitivity of the ML predictions increases.

Table 3
Post-SMOTE decision tree performance on the testing sample (variable: WC crime rate).

		Real status		
Year: 2012		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	**WC crime rate = 0**	2024	60	2084
	WC crime rate = 1	320	173	493
	Total	2344	233	2577
	Correctly predicted	**86.4 %**	**74.3 %**	**85.3 %**
Year: 2013		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	**WC crime rate = 0**	2018	66	2084
	WC crime rate = 1	320	173	493
	Total	2338	239	2577
	Correctly predicted	**86.3 %**	**72.4 %**	**85 %**
Year: 2014		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	**WC crime rate = 0**	2013	71	2084
	WC crime rate = 1	309	184	493
	Total	2322	255	2577
	Correctly predicted	**86.7 %**	**72.2 %**	**85.3 %**

Notes: Out-of-sample estimation on the testing subsample, using the model trained on the rebalanced training subsample. WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) is positive and 0 otherwise.

Table 4
Post-SMOTE decision tree performance on the testing sample (variable: Δ WC crime rate).

		Real status		
Year: 2012		**Δ WC crime rate = 0**	**Δ WC crime rate = 1**	**Total**
Predicted status	**Δ WC crime rate =0**	1831	36	1867
	Δ WC crime rate = 1	566	144	710
	Total	2397	180	2577
	Correctly predicted	**76.4 %**	**80 %**	**76.6 %**
Year: 2013		**Δ WC crime rate = 0**	**Δ WC crime rate = 1**	**Total**
Predicted status	**Δ WC crime rate =0**	1838	29	1867
	Δ WC crime rate = 1	627	83	710
	Total	2465	112	2577
	Correctly predicted	**74.6 %**	**74.1 %**	**74.5 %**
Year: 2014		**Δ WC crime rate = 0**	**Δ WC crime rate = 1**	**Total**
Predicted status	**Δ WC crime rate = 0**	1830	37	1867
	Δ WC crime rate = 1	567	143	710
	Total	2397	180	2577
	Correctly predicted	**76.4 %**	**79.4 %**	**76.6 %**

Notes: Out-of-sample estimation on the testing subsample, using the model trained on the rebalanced training subsample. Δ WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) has increased with respect to the previous year, and 0 otherwise.

Table 5
Anti-corruption threshold vs decision tree rule performances (testing sample; variable: WC crime rate).

		Real status		
Year: 2012		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	Correctly predicted	86.4 %	74.3 %	85.3 %
Anti-corruption threshold	Correctly targeted	95.8 %	48.9 %	91.5 %
	Difference	**− 9.4 %**	**+ 25.4 %**	**− 6.2 %**
Year: 2013		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	Correctly predicted	86.3 %	72.4 %	85 %
Anti-corruption threshold	Correctly targeted	95.6 %	46 %	91 %
	Difference	**− 9.3 %**	**+ 26.4 %**	**− 6 %**
Year: 2014		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	Correctly predicted	86.7 %	72.2 %	85.3 %
Anti-corruption threshold	Correctly targeted	95.4 %	41.6 %	90.1 %
	Difference	**− 8.7 %**	**+ 30.6 %**	**− 4.8 %**
Years: 2012 - 2014		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Overall average difference in performance		**− 9.1 %**	**+ 27.5 %**	**− 5.7 %**

Notes: The comparison is on the 2577 municipalities belonging to the testing subsample. WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) is positive and 0 otherwise.

6. Algorithmic forecasts in the service of the anti-corruption law

In 2012 Law 190,[14] named "Rules for the prevention and repression of corruption and unlawfulness in public administration" and informally known as "Legge Severino" (from the name of the then Minister of Justice) introduced new and more stringent criteria to fight corruption in Italy. For instance, it expanded the definition of corruption and enhanced transparency and disclosure requirements for public-sector workers. On top of these general prescriptions, which apply to the entire Italian public administration, the law also introduced a number of additional restrictions related to the possibility of assigning directive positions in public administrations to those who had held political responsibilities in the previous years. Crucially, at the local level, these more restrictive rules only apply to municipalities with more than 15 thousand inhabitants.[15] The rationale behind excluding smaller municipalities is that the costs related to the regulation are presumably larger than the associated benefits in smaller municipalities as they have very few cases of white-collar crime episodes anyway. Smaller municipalities also receive fewer public resources, which makes them, in principle, less exposed to corruption risk. According to new estimates provided by De Angelis et al. (2020), Law 190/2012 seems to have had a positive impact, at the least in the South of Italy, where municipalities with over 15 thousand residents experienced fewer corruption episodes linked to EU regional transfers after 2012.

We show now that our algorithms can help strengthen effectiveness of the anti-corruption regulation. We consider (as before) the set of municipalities belonging to the training set (2577) and evaluate the accuracy of the ML algorithm and the anti-corruption threshold when predicting (a) municipalities with a positive crime rate and (b) municipalities with an increasing crime rate. Table 5 considers the outcome in levels.[16] Note that the threshold envisaged by the law does an excellent job as for the $y = 0$. For instance, in 2012, municipalities under the cutoff that do not experience corruption episodes represent 95.8 % of the sample. Conversely, above the cutoff, only 48.9 % of the municipalities are caught in the more severe anti-corruption net, while the others are missed by the law. ML predictions imply a huge gain in sensitivity (+27.5 % over the three years) at the cost of a reduced specificity (−9.1 %). Table 6 provides the same analysis for the outcome defined in variations. A similar pattern emerges. With the ML forecast, sensitivity would have risen by 43.6 % over the entire time span considered; on the other hand, specificity would have been reduced by 17.7 %.

7. Open issues and conclusions

We have documented that the gains from using ML predictive tools are substantial. It is worth noting that these gains might well represent a conservative estimate. First, we chose our ML algorithm on the basis of its transparency. More complex - but admittedly less transparent - algorithms might provide better performances. In this regard, we also employed random forest algorithms, but we did not get significant

[14] See http://www.anticorruzione.it/portal/public/classic/MenuServizio/FAQ/Anticorruzione.

[15] Cf. Articles 7, 8, 11, 12, 13, 14 of Legislative Decree no. 39/2013.

[16] A potential issue refers to the timing of the introduction of the law. To the extent that the entry into force of the law affects levels and trends of corruption, prediction accuracy might be lower as our algorithm is trained on pre-intervention data (the law came into effect only in 2013). However, the results for 2013 and 2014 are very similar to those obtained for 2012.

Table 6
Anti-corruption threshold vs decision tree rule performances (testing sample; variable: Δ WC crime rate).

		Real status		
Year: 2012		**Δ WC crime rate = 0**	**Δ WC crime rate = 1**	**Total**
Predicted status	Correctly predicted	76.4 %	80 %	76.6 %
Anti-corruption threshold	Correctly targeted	94.5 %	45.6 %	91.1 %
	Difference	**− 18.1 %**	**+ 34.4 %**	**− 14.5 %**
Year: 2013		**Δ WC crime rate = 0**	**Δ WC crime rate = 1**	**Total**
Predicted status	Correctly predicted	74.6 %	74.1 %	74.5 %
Anti-corruption threshold	Correctly targeted	92.3 %	21.4 %	89.3 %
	Difference	**− 17.7 %**	**+ 52.7 %**	**− 14.8 %**
Year: 2014		**Δ WC crime rate = 0**	**Δ WC crime rate = 1**	**Total**
Predicted status	Correctly predicted	76.4 %	79.4 %	76.6 %
Anti-corruption threshold	Correctly targeted	93.8 %	35.6 %	89.7 %
	Difference	**− 17.4 %**	**+ 43.8 %**	**− 13.1 %**
Years: 2012 - 2014		**Δ WC crime rate = 0**	**Δ WC crime rate = 1**	**Total**
Overall average difference in performance		**− 17.7 %**	**+ 43.6 %**	**− 14.1 %**

Notes: The comparison is on the 2577 municipalities belonging to the testing subsample. Δ WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) has increased with respect to the previous year, and 0 otherwise.

improvements over the accuracy of the classification trees.[17]

When compared with Law 190/2012, we have been able to improve predictive sensitivity (y = 1) at the expense of losing specificity (y = 0). Given the high socio-economic costs of corruption, the benefits related to higher sensitivity seem to be warranted. Nonetheless, the pool of municipalities predicted as potentially 'corrupted' is higher under ML than the law cutoff (according to Table 5, in 2012, this pool includes 493 municipalities versus 213 of them). This means that if ML predictions are taken as to select places where to implement stricter anti-corruption rules, as an alternative to the cutoff envisaged by Law 190/2012, then the overall regulatory costs would, in principle, rise. In any case, ML forecasts could be adopted to prioritize anti-corruption efforts on the ground, such as those related to police investigations, excluding municipalities above 15,000 inhabitants that are predicted not to experience an increase in corruption episodes. Under this scenario, overall regulatory costs will be lower. For instance, in the year 2012, 10.8 % and 10.3 % of cities with more than 15,000 inhabitants, respectively for the outcomes in levels and variations, can be sheltered from the new rules notwithstanding they are above the demographic threshold.

As for the outcomes, the ones we have proposed are admittedly the simplest ones when it comes to figuring out the preferences of the policymaker. They can be usefully combined. For instance, a policymaker might be particularly worried about places where both dummies (levels and variations) take the value of one. On the other hand, it could be that repression efforts have to be concentrated at the early stages of a corruption escalation, so the policymaker might care more about places that move from zero to positive corruption. Again, the authorities might choose to focus on the municipalities in which the number of corruption episodes reaches a given threshold. These, and other similar cases, are easily accommodated in our framework.

An important focus in the ML literature is its potential bias. Suppose that our data are contaminated because corruption episodes are more likely to be reported in certain communities, for instance in municipalities with higher social capital (Putnam et al., 1992). If this is the case, then the ML prediction is likely to be biased as well, and the municipalities with higher endowments of social capital are most likely to be classified as experiencing an increase in corruption, ceteris paribus. Contamination issues have no sensible solution. If the y = 0 are false negatives, because in those municipalities there are white-collar crimes not recorded in the SDI, there is little to do. However, our artificially rebalanced training sample (the one on which the prediction is based) is likely to be less exposed to such a bias, compared to the original skewed sample. SMOTE undersamples the most numerous class by excluding those observations which are less similar – as measured by comparing observable features - to the other class. Therefore, the post-SMOTE sample is likely to be featured by high similarity in observables and – following the Altonji et al. (2005)'s argument – in unobservables as well. In any case, contamination issues have no easy solution. Kleinberg et al. (2020) suggest that it is the use of data-driven methods, rather than non-quantitative approaches, which ensures more progress on this front.

We have picked up classification trees purposely to minimize transparency issues. Our classification tree is intuitive and easy to understand even without a strong statistical background. This makes them appealing for policy-targeting purposes in a hypothetical scenario in which it was possible to use an algorithm to decide on the areas in which to apply some prescriptions of law. Obviously, a mechanism based just on one single threshold, like the one envisaged by Law 190/2012, is easier to understand. However, the cost of endorsing a tree rather than a single population-based threshold rule might be considered not that large: forecasts based on the decision tree increase effectiveness in detecting 'corrupted' municipalities, while the population threshold

[17] Results are available upon request.

does not have such a sound foundation. The increase in complexity can hence be justified and communicated as necessary to serve a public aim. Another aspect refers to the amount of information that the policymaker needs. Algorithm predictive power increases with more information. However, we have shown that, even with a not impressive list of features, gains are substantial. Concerning information requirements, also note that the trees have the advantage of using very few variables once its structure has been defined, which is after the phase of training and testing (Section 3). In our case, predictions need only 3 (Fig. 3) or 6 (Fig. 4) variables. This is not the case for more complex algorithms, which require the whole information set at any step. Also related to transparency, ML methods can highlight the targeting that an authority interested in fighting corruption should adopt. Therefore, they can also provide information on whether other additional objectives, such as omitted payoffs (Kleinberg et al., 2018), have a role in this important kind of public decisions. For instance, corrupt politicians might conspire to order police investigations far from certain places. Having the ML prediction map, which can be easily compared to areas with actual police efforts, might shed light on such episodes.

For ML-based predictive policing to be effective, the capabilities to access both the technology and the necessary data must be adequate. Concerning such aspects, today, a wide and growing array of data, more and more granular, both structured (such as individual crime records) and unstructured (such as cctv and security camera footage), is theoretically available for predictive policing in many countries. However, the extent of data availability is arguably heterogeneous across territories, with larger cities providing a much greater amount of information (think, for instance, at cctv footage). Hence, in order for predictive policing to be more effective, it must be accompanied by processes of data production and consolidation that aims to fill such gaps. Relatedly, since the predictions rely on past data, the data must be constantly updated in order to avoid biased predictions.

In order to tackle these challenges, predictive policy research must be supported by advantages in information systems research. In particular, the ability to exploit very high-frequency data (possibly at ZIP code level) and individuate possible predictors from unconventional data sources, such as raster data and text data from social networks, will be key. For example, raster data can help to identify those locations potentially subject to burglaries (see, for instance, Groff and La Vigne, 2001) or to white-collar crimes, such as new building sites. At the same time, the ability to gather and process real-time data would reduce the risk that predictive policing is based on biased predictions. Concerning real-time data, the technological applications involve, for instance, Automated Facial Recognition and voice identification systems (Jansen, 2018).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Appendix A. Descriptive statistics and additional results

Table A.1
Full list of features employed.

Variable	Year	Source
Corruption crime rate in the previous year	2011	Italian Ministry of the Interior – *Sistema d'Indagine*
Lag 1of the outcome variable	2011	Italian Ministry of the Interior – *Sistema d'Indagine*
Lag 2 of the outcome variable	2010	Italian Ministry of the Interior – *Sistema d'Indagine*
Lag 3 of the outcome variable	2009	Italian Ministry of the Interior – *Sistema d'Indagine*
Lag 4 of the outcome variable	2008	Italian Ministry of the Interior – *Sistema d'Indagine*
Population	2011	Italian National Institute of Statistics (Istat)
Annual intercensal change in population	2011	Italian National Institute of Statistics (Istat)
Annual intercensal change in population under 15 years old	2011	Italian National Institute of Statistics (Istat)
Annual intercensal change in population above 15 years old	2011	Italian National Institute of Statistics (Istat)
Municipality surface covered by residential areas	2011	Italian National Institute of Statistics (Istat)
Share of population living in residential areas	2011	Italian National Institute of Statistics (Istat)
Population density	2011	Italian National Institute of Statistics (Istat)
Ratio between male and female population	2011	Italian National Institute of Statistics (Istat)
Share of population under 6 years old	2011	Italian National Institute of Statistics (Istat)
Share of population above 75 years old	2011	Italian National Institute of Statistics (Istat)
Ratio between population above 65 years old and working-age population	2011	Italian National Institute of Statistics (Istat)
Ratio between population below 15 years old and working-age population	2011	Italian National Institute of Statistics (Istat)
Ratio between population above 65 years old and population below 15 years old	2011	Italian National Institute of Statistics (Istat)
Ratio between legally divorced population and population above 17 years old	2011	Italian National Institute of Statistics (Istat)
Number of foreign people per 1000 inhabitants	2011	Italian National Institute of Statistics (Istat)
Ratio between foreign population below 18 years old and total foreign population	2011	Italian National Institute of Statistics (Istat)
Ratio between married couples with a foreign spouse and total number of married couples	2011	Italian National Institute of Statistics (Istat)
Employment rate of the foreign population	2011	Italian National Institute of Statistics (Istat)
Ratio between the employment rate of the Italian population and that of the foreign population	2011	Italian National Institute of Statistics (Istat)
Ratio between the unemployment rate of the Italian population and that of the foreign population	2011	Italian National Institute of Statistics (Istat)
Index of residential mobility of foreign population	2011	Italian National Institute of Statistics (Istat)
Index of school attendance of foreign population	2011	Italian National Institute of Statistics (Istat)
Ratio between the index school attendance of the Italian population and that of the foreign population	2011	Italian National Institute of Statistics (Istat)
Ratio between the share of Italian independent workers and the share of foreign independent workers	2011	Italian National Institute of Statistics (Istat)
Average household size	2011	Italian National Institute of Statistics (Istat)
Share of one-person households	2011	Italian National Institute of Statistics (Istat)

(*continued on next page*)

Table A.1 (*continued*)

Variable	Year	Source
Share of households with two or more family nuclei	2011	Italian National Institute of Statistics (Istat)
Share of young people (below 35 years old) living alone	2011	Italian National Institute of Statistics (Istat)
Share of young single-parent households	2011	Italian National Institute of Statistics (Istat)
Share of young couples without children	2011	Italian National Institute of Statistics (Istat)
Share of young couples with children	2011	Italian National Institute of Statistics (Istat)
Share of old single-parent households	2011	Italian National Institute of Statistics (Istat)
Share of old couples without children	2011	Italian National Institute of Statistics (Istat)
Share of old couples with children	2011	Italian National Institute of Statistics (Istat)
Share of real estate ownership among households	2011	Italian National Institute of Statistics (Istat)
Mean surface of inhabited buildings (sq. m.)	2011	Italian National Institute of Statistics (Istat)
Share of buildings into disuse	2011	Italian National Institute of Statistics (Istat)
Share of buildings into disuse in residential areas	2011	Italian National Institute of Statistics (Istat)
Share of buildings into disuse in non-residential areas	2011	Italian National Institute of Statistics (Istat)
Average building age	2011	Italian National Institute of Statistics (Istat)
Service availability index	2011	Italian National Institute of Statistics (Istat)
Share of well-preserved buildings	2011	Italian National Institute of Statistics (Istat)
Share of uninhabitable buildings	2011	Italian National Institute of Statistics (Istat)
Share of historical buildings currently inhabited	2011	Italian National Institute of Statistics (Istat)
Index of building growth	2011	Italian National Institute of Statistics (Istat)
Average number of square meters per inhabitant	2011	Italian National Institute of Statistics (Istat)
Share of undercrowded buildings	2011	Italian National Institute of Statistics (Istat)
Share of overcrowded buildings	2011	Italian National Institute of Statistics (Istat)
Index of residential mobility	2011	Italian National Institute of Statistics (Istat)
Index of gender differences in high school and tertiary education	2011	Italian National Institute of Statistics (Istat)
Share of adults attending learning courses	2011	Italian National Institute of Statistics (Istat)
Share of young people with a university degree	2011	Italian National Institute of Statistics (Istat)
Ratio between the share of adult people with high school or university education and the share of adult people with middle school education	2011	Italian National Institute of Statistics (Istat)
Share of illiterate people	2011	Italian National Institute of Statistics (Istat)
Share of young people leaving school early	2011	Italian National Institute of Statistics (Istat)
Share of adults with high-school or tertiary education	2011	Italian National Institute of Statistics (Istat)
Share of young people with tertiary education	2011	Italian National Institute of Statistics (Istat)
Education level of people between 15 and 19 years old	2011	Italian National Institute of Statistics (Istat)
Share of adults with middle-school education	2011	Italian National Institute of Statistics (Istat)
Ratio between active and non-active young people	2011	Italian National Institute of Statistics (Istat)
Male unemployment rate	2011	Italian National Institute of Statistics (Istat)
Female unemployment rate	2011	Italian National Institute of Statistics (Istat)
Unemployment rate	2011	Italian National Institute of Statistics (Istat)
Youth unemployment rate	2011	Italian National Institute of Statistics (Istat)
Daily mobility inside the municipality for study or work (share of the working-age population)	2011	Italian National Institute of Statistics (Istat)
Daily mobility outside the municipality for study or work (share of the working-age population)	2011	Italian National Institute of Statistics (Istat)
Daily worker mobility outside the municipality	2011	Italian National Institute of Statistics (Istat)
Daily student mobility outside the municipality	2011	Italian National Institute of Statistics (Istat)
Daily mobility with private transport	2011	Italian National Institute of Statistics (Istat)
Daily mobility with public transport	2011	Italian National Institute of Statistics (Istat)
Daily walking and cycling mobility	2011	Italian National Institute of Statistics (Istat)
Daily mobility for short distances	2011	Italian National Institute of Statistics (Istat)
Daily mobility for long distances	2011	Italian National Institute of Statistics (Istat)
Vulnerability index	2011	Italian National Institute of Statistics (Istat)
Place in vulnerability index ranking	2011	Italian National Institute of Statistics (Istat)
Share of provincial population living in very vulnerable municipalities	2011	Italian National Institute of Statistics (Istat)
Share of people living in makeshift accommodation	2011	Italian National Institute of Statistics (Istat)
Share of households with more than 5 members	2011	Italian National Institute of Statistics (Istat)
Share of households in potential economic hardships	2011	Italian National Institute of Statistics (Istat)
Share of population living in overcrowded homes	2011	Italian National Institute of Statistics (Istat)
Share of young adults outside the job market	2011	Italian National Institute of Statistics (Istat)
Share of households needing health assistance	2011	Italian National Institute of Statistics (Istat)
Share of foreign people from Eastern Europe	2011	Italian National Institute of Statistics (Istat)
Share of foreign people from Northern Africa	2011	Italian National Institute of Statistics (Istat)
Share of foreign people from Southern Europe	2011	Italian National Institute of Statistics (Istat)
Absolute number of foreign people	2011	Italian National Institute of Statistics (Istat)
Number of police stations within 50 km from the municipality centroid	2011	DatiOpen.it
Number of kilometers from the municipality centroid to the nearest police station	2011	DatiOpen.it
Average temperature (°C)	2011	

(*continued on next page*)

Table A.1 (*continued*)

Variable	Year	Source
		University of Delaware (Matsuura and Willmott, 2015)
Total precipitation (mm)	2011	University of Delaware (Matsuura and Willmott, 2015)

Notes: The same set of features is used to carry out the two predictive tasks, with the exception of the lagged outcome variables, for which only lags of the corresponding outcome are used.

Table A.2
The Accuracy Paradox: pre-SMOTE decision tree performance on the testing sample (variable: WC crime rate; year: 2012).

		Real status		
		WC crime rate = 0	**WC crime rate = 1**	**Total**
Predicted status	**WC crime rate = 0**	2311	159	2470
	WC crime rate = 1	33	74	107
	Total	2344	233	2577
	Correctly predicted	**98.6 %**	**31.2 %**	**92.6 %**

Notes: Out-of-sample estimation on the testing subsample, using the model trained on the original imbalanced training subsample. WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) is positive and 0 otherwise.

Table A.3
The Accuracy Paradox: pre-SMOTE decision tree performance on the testing sample (variable: Δ WC crime rate; year: 2012).

		Real status		
		Δ WC crime rate = 0	**Δ WC crime rate = 1**	**Total**
Predicted status	**Δ WC crime rate = 0**	2397	180	2577
	Δ WC crime rate = 1	0	0	0
	Total	2397	180	2577
	Correctly predicted	**100 %**	**0 %**	**93 %**

Notes: Out-of-sample estimation on the testing subsample, using the model trained on the original imbalanced training subsample. Δ WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) has increased with respect to the previous year, and 0 otherwise.

Table A.4
Post-SMOTE Logit performance on the testing sample (variable: WC crime rate).

		Real status		
Year: 2012		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	**WC crime rate = 0**	1972	65	2037
	WC crime rate = 1	372	168	540
	Total	2344	233	2577
	Correctly predicted	**84.1 %**	**72.1 %**	**83 %**
Year: 2013		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	**WC crime rate = 0**	1972	65	2037
	WC crime rate = 1	366	174	540
	Total	2338	239	2577
	Correctly predicted	**84.4 %**	**72.8 %**	**83.3 %**
Year: 2014		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	**WC crime rate = 0**	1962	75	2037
	WC crime rate = 1	360	180	540
	Total	2322	255	2577
	Correctly predicted	**84.5 %**	**70.6 %**	**83.1 %**

Notes: Out-of-sample estimation on the testing subsample, using the model trained on the rebalanced training subsample. WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) is positive and 0 otherwise. In order to allow comparability with the decision tree, missing values in the testing sample have been imputed using the *rfImpute* package in *R*.

Table A.5
Post-SMOTE Logit performance on the testing sample (variable: Δ WC crime rate).

		Real status		
Year: 2012		**Δ WC crime rate = 0**	**Δ WC crime rate = 1**	**Total**
Predicted status	**Δ WC crime rate =0**	1906	61	1967
	Δ WC crime rate = 1	491	119	610
	Total	2397	180	2577
	Correctly predicted	**79.5 %**	**66.1 %**	**78.6 %**
Year: 2013		**Δ WC crime rate = 0**	**Δ WC crime rate = 1**	**Total**
Predicted status	**Δ WC crime rate =0**	1919	48	1967
	Δ WC crime rate = 1	546	64	610
	Total	2465	112	2577
	Correctly predicted	**77.9 %**	**57.1 %**	**77 %**
Year: 2014		**Δ WC crime rate = 0**	**Δ WC crime rate = 1**	**Total**
Predicted status	**Δ WC crime rate = 0**	1896	71	1967
	Δ WC crime rate = 1	501	109	610
	Total	2397	180	2577
	Correctly predicted	**79.1 %**	**60.6 %**	**77.8 %**

Notes: Out-of-sample estimation on the testing subsample, using the model trained on the rebalanced training subsample. Δ WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) has increased with respect to the previous year, and 0 otherwise. In order to allow comparability with the decision tree, missing values in the testing sample have been imputed using the *rfImpute* package in *R*.

Variable	Year	Source
Corruption crime rate in the previous year	2011	Italian Ministry of the Interior – *Sistema d'Indagine*
Lag 1of the outcome variable	2011	Italian Ministry of the Interior – *Sistema d'Indagine*
Lag 2 of the outcome variable	2010	Italian Ministry of the Interior – *Sistema d'Indagine*
Lag 3 of the outcome variable	2009	Italian Ministry of the Interior – *Sistema d'Indagine*
Lag 4 of the outcome variable	2008	Italian Ministry of the Interior – *Sistema d'Indagine*
Population	2011	Italian National Institute of Statistics (Istat)
Annual intercensal change in population	2011	Italian National Institute of Statistics (Istat)
Annual intercensal change in population under 15 years old	2011	Italian National Institute of Statistics (Istat)
Annual intercensal change in population above 15 years old	2011	Italian National Institute of Statistics (Istat)
Municipality surface covered by residential areas	2011	Italian National Institute of Statistics (Istat)
Share of population living in residential areas	2011	Italian National Institute of Statistics (Istat)
Population density	2011	Italian National Institute of Statistics (Istat)
Ratio between male and female population	2011	Italian National Institute of Statistics (Istat)
Share of population under 6 years old	2011	Italian National Institute of Statistics (Istat)
Share of population above 75 years old	2011	Italian National Institute of Statistics (Istat)
Ratio between population above 65 years old and working-age population	2011	Italian National Institute of Statistics (Istat)
Ratio between population below 15 years old and working-age population	2011	Italian National Institute of Statistics (Istat)
Ratio between population above 65 years old and population below 15 years old	2011	Italian National Institute of Statistics (Istat)
Ratio between legally divorced population and population	2011	Italian National Institute of Statistics (Istat)

Fig. A.1. Classification tree for WC crime rate – Pre-SMOTE data.

above 17 years old		
Number of foreign people per 1000 inhabitants	2011	Italian National Institute of Statistics (Istat)
Ratio between foreign population below 18 years old and total foreign population	2011	Italian National Institute of Statistics (Istat)
Ratio between married couples with a foreign spouse and total number of married couples	2011	Italian National Institute of Statistics (Istat)
Employment rate of the foreign population	2011	Italian National Institute of Statistics (Istat)
Ratio between the employment rate of the Italian population and that of the foreign population	2011	Italian National Institute of Statistics (Istat)
Ratio between the unemployment rate of the Italian population and that of the foreign population	2011	Italian National Institute of Statistics (Istat)
Index of residential mobility of foreign population	2011	Italian National Institute of Statistics (Istat)
Index of school attendance of foreign population	2011	Italian National Institute of Statistics (Istat)
Ratio between the index school attendance of the Italian population and that of the foreign population	2011	Italian National Institute of Statistics (Istat)
Ratio between the share of Italian independent workers and the share of foreign independent workers	2011	Italian National Institute of Statistics (Istat)
Average household size	2011	Italian National Institute of Statistics (Istat)
Share of one-person households	2011	Italian National Institute of Statistics (Istat)
Share of households with two or more family nuclei	2011	Italian National Institute of Statistics (Istat)
Share of young people (below 35 years old) living alone	2011	Italian National Institute of Statistics (Istat)
Share of young single-parent households	2011	Italian National Institute of Statistics (Istat)
Share of young couples without children	2011	Italian National Institute of Statistics (Istat)
Share of young couples with children	2011	Italian National Institute of Statistics (Istat)
Share of old single-parent households	2011	Italian National Institute of Statistics (Istat)
Share of old couples without children	2011	Italian National Institute of Statistics (Istat)
Share of old couples with children	2011	Italian National Institute of Statistics (Istat)
Share of real estate ownership among households	2011	Italian National Institute of Statistics (Istat)

Fig. A.1. (*continued*).

Mean surface of inhabited buildings (sq. m.)	2011	Italian National Institute of Statistics (Istat)
Share of buildings into disuse	2011	Italian National Institute of Statistics (Istat)
Share of buildings into disuse in residential areas	2011	Italian National Institute of Statistics (Istat)
Share of buildings into disuse in non-residential areas	2011	Italian National Institute of Statistics (Istat)
Average building age	2011	Italian National Institute of Statistics (Istat)
Service availability index	2011	Italian National Institute of Statistics (Istat)
Share of well-preserved buildings	2011	Italian National Institute of Statistics (Istat)
Share of uninhabitable buildings	2011	Italian National Institute of Statistics (Istat)
Share of historical buildings currently inhabited	2011	Italian National Institute of Statistics (Istat)
Index of building growth	2011	Italian National Institute of Statistics (Istat)
Average number of square meters per inhabitant	2011	Italian National Institute of Statistics (Istat)
Share of undercrowded buildings	2011	Italian National Institute of Statistics (Istat)
Share of overcrowded buildings	2011	Italian National Institute of Statistics (Istat)
Index of residential mobility	2011	Italian National Institute of Statistics (Istat)
Index of gender differences in high school and tertiary education	2011	Italian National Institute of Statistics (Istat)
Share of adults attending learning courses	2011	Italian National Institute of Statistics (Istat)
Share of young people with a university degree	2011	Italian National Institute of Statistics (Istat)
Ratio between the share of adult people with high school or university education and the share of adult people with middle school education	2011	Italian National Institute of Statistics (Istat)
Share of illiterate people	2011	Italian National Institute of Statistics (Istat)
Share of young people leaving school early	2011	Italian National Institute of Statistics (Istat)
Share of adults with high-school or tertiary education	2011	Italian National Institute of Statistics (Istat)
Share of young people with tertiary education	2011	Italian National Institute of Statistics (Istat)
Education level of people between 15 and 19 years old	2011	Italian National Institute of Statistics (Istat)
Share of adults with middle-school education	2011	Italian National Institute of Statistics (Istat)

Fig. A.1. (*continued*).

Ratio between active and non-active young people	2011	Italian National Institute of Statistics (Istat)
Male unemployment rate	2011	Italian National Institute of Statistics (Istat)
Female unemployment rate	2011	Italian National Institute of Statistics (Istat)
Unemployment rate	2011	Italian National Institute of Statistics (Istat)
Youth unemployment rate	2011	Italian National Institute of Statistics (Istat)
Daily mobility inside the municipality for study or work (share of the working-age population)	2011	Italian National Institute of Statistics (Istat)
Daily mobility outside the municipality for study or work (share of the working-age population)	2011	Italian National Institute of Statistics (Istat)
Daily worker mobility outside the municipality	2011	Italian National Institute of Statistics (Istat)
Daily student mobility outside the municipality	2011	Italian National Institute of Statistics (Istat)
Daily mobility with private transport	2011	Italian National Institute of Statistics (Istat)
Daily mobility with public transport	2011	Italian National Institute of Statistics (Istat)
Daily walking and cycling mobility	2011	Italian National Institute of Statistics (Istat)
Daily mobility for short distances	2011	Italian National Institute of Statistics (Istat)
Daily mobility for long distances	2011	Italian National Institute of Statistics (Istat)
Vulnerability index	2011	Italian National Institute of Statistics (Istat)
Place in vulnerability index ranking	2011	Italian National Institute of Statistics (Istat)
Share of provincial population living in very vulnerable municipalities	2011	Italian National Institute of Statistics (Istat)
Share of people living in makeshift accommodation	2011	Italian National Institute of Statistics (Istat)
Share of households with more than 5 members	2011	Italian National Institute of Statistics (Istat)
Share of households in potential economic hardships	2011	Italian National Institute of Statistics (Istat)
Share of population living in overcrowded homes	2011	Italian National Institute of Statistics (Istat)
Share of young adults outside the job market	2011	Italian National Institute of Statistics (Istat)
Share of households needing health assistance	2011	Italian National Institute of Statistics (Istat)
Share of foreign people from Eastern Europe	2011	Italian National Institute of Statistics (Istat)
Share of foreign people from Northern Africa	2011	Italian National Institute of Statistics (Istat)
Share of foreign people from Southern Europe	2011	Italian National Institute of Statistics (Istat)
Absolute number of foreign people	2011	Italian National Institute of Statistics (Istat)
Number of police stations within 50 km from the municipality centroid	2011	DatiOpen.it
Number of kilometers from the municipality centroid to the nearest police station	2011	DatiOpen.it
Average temperature (°C)	2011	University of Delaware (Matsuura & Willmott, 2015)
Total precipitation (mm)	2011	University of Delaware (Matsuura & Willmott, 2015)

Fig. A.1. (*continued*).

Table A.6
Post-SMOTE Decision tree performance on the testing sample without lagged crime predictors (variable: WC crime rate).

		Real status		
Year: 2012		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	**WC crime rate = 0**	1886	55	1941
	WC crime rate = 1	458	178	636
	Total	2344	233	2577
	Correctly predicted	**80.5 %**	**76.4 %**	**80.1 %**
Year: 2013		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	**WC crime rate = 0**	1893	48	1941
	WC crime rate = 1	445	191	636
	Total	2338	239	2577
	Correctly predicted	**81 %**	**79.9 %**	**80.9 %**
Year: 2014		**WC crime rate = 0**	**WC crime rate = 1**	**Total**
Predicted status	**WC crime rate = 0**	1886	55	1941
	WC crime rate = 1	436	200	636
	Total	2322	255	2577
	Correctly predicted	**81.2 %**	**78.4 %**	**81 %**

Notes: Out-of-sample estimation on the testing subsample, using the model trained on the rebalanced training subsample. WC crime rate is a binary variable taking value 1 if the white-collar crime rate (i.e., the number of white-collar crimes per 1000 inhabitants) is positive and 0 otherwise.

Appendix B. Methodological annex

B.1. Machine learning

Machine learning is a subfield of artificial intelligence. ML algorithms have been developed in computer science and statistical literature to deal with prediction problems (Varian, 2014). *Supervised* ML involves building a statistical model for predicting an outcome of interest (*output*) based on a set of features (*inputs*) (Lantz, 2019). There is a training set of data in which both the outcome and features can be observed. Using this data, a prediction model is built to predict the outcome for new unseen objects. A good model is one that accurately predicts such an outcome and minimizes the predictive error on previously unseen data (the so-called 'test error') (Hastie et al., 2009). The standard ML routine thus consists of randomly splitting the original sample into two separate subsets, a training set, and a testing (hold-out) sample. The split has to be necessarily random so to avoid including systematic differences between the two separate sets (Lantz, 2019). Also, the predictive model must be evaluated on the testing set, because measures of performance evaluated on the training sample tend to be overoptimistic concerning the true model performance, as the algorithm is evaluating the model on data it already 'knows'. The best practice is thus to assess a model's performance on data it has never seen before, as this introduces a 'firewall' principle: none of the data involved in generating the prediction function is used to evaluate it (Mullainathan and Spiess, 2017). As for the random split, we employ, as specified in Section 3.1, the conventional choice of using 2/3 of the dataset as the training set and the remaining 1/3 of observations as the testing set, which is among the most popular splitting rules (Zhao and Chen, 2013). We then use the training subset to train and tune our algorithms and the testing subset to estimate their future performance. The out-of-sample performance of the model on the held-out data constitutes a reliable and generalizable measure of its actual predictive ability on future data.

B.2. Classification tree

A classification tree is a supervised machine learning technique that is based on the concept of recursive partitioning, also known as the 'divide and conquer' approach (Lantz, 2019). Through the iterative process of recursive binary splitting, the algorithm 'grows' a tree by repeatedly splitting the data into smaller and smaller subsets to identify important patterns that can be employed for predicting a qualitative outcome. The process goes on until sufficient within-subset homogeneity (or another stopping criterion) is reached. This makes it an extremely flexible method that can automatically detect nonlinearities and interactions among predictors through the sequence of splits. Trees are known to suffer from high variance, i.e., they are quite sensitive to small changes in the training sample and tend to be prone to overfitting. Albeit one could grow a very complex tree, large enough so that no observation is misclassified, a high number of levels in a tree is very likely to result in high variance and to overfit the data, leading to a predictive model with poor out-of-sample performance. This is why regularization, via the so-called 'pruning', is used to tune the algorithm and prevent the risk of in-sample overfitting. Pruning means setting a penalization cost for flexibility; this cost takes the name of 'complexity parameter' (*cp*). Setting a low *cp* would lead to a large tree with a good fit in the training sample, but possibly with a large out-of-sample error. By setting a higher *cp*, we reduce its size (we "prune" the tree) and reduce the risk of overfitting. In order to select the optimal value of *cp*, which maximizes the out-of-sample accuracy of our model, we employ 10-fold cross-validation for model selection in the training dataset, compare the ten different cross-

validation errors, and pick up the complexity parameter associated with the lowest error for the final model used to predict observations belonging to the testing set. Note that this hyperparameter optimization via cross-validation is only done on the training set.

B.3. SMOTE

In order to tackle the empirical challenge stemming from our highly imbalanced data, we employ a well-known solution: rebalancing the training set. Specifically, we leverage the Synthetic Minority Oversampling Technique (SMOTE) routine, proposed by Chawla et al. (2002), to rebalance the two classes in our training sample. SMOTE is a popular rebalancing algorithm that oversamples the under-represented cases and undersamples the majority class, leading to a smaller rebalanced dataset. More in detail, SMOTE oversamples the minority class by introducing synthetic examples along the line segments joining any/all of the *k* minority class nearest neighbors. Depending on the amount of over-sampling required, neighbors from the *k* nearest neighbors are then randomly chosen to generate the synthetic observations.

We adopt the common *k* value of ten nearest neighbors and the standard amount of oversampling equal to 200 %. Importantly, as specified in Section 3.1, we implement the SMOTE algorithm only on the training subsample, i.e., the set on which when train our model, leaving the testing sample, i.e., the set on which we evaluate its future performance, untouched. Therefore, the training dataset is artificially balanced over the two outcomes, while the prediction is tested on the original skewed sample (i.e., on real-world corruption crime data). Without applying SMOTE on the training data, the classification tree algorithm would simply predict the majority class (Y = 0), resulting in predictive performances with a very high overall accuracy but low or null sensitivity.

B.4. Confusion matrix

After training and tuning the algorithm on the training sample, we evaluate its out-of-sample performance in the testing set via confusion matrices through which we compare the predicted and actual values of our binary outcome. A confusion matrix is a widely known analytical tool that is employed to evaluate predictive models' performances for classification tasks. In the case of classification problems with a binary outcome, it consists of a simple two-way table such as the one reported below:

Table B.1
Example of confusion matrix for a binary classification problem.

		Real status	
		Y = 0	Y = 1
Predicted status	Y = 0	True negatives	False negatives
	Y = 1	False positives	True positives

The *True negatives* cell contains negative cases (Y = 0) that were correctly identified. The *True positives* cell includes the positive instances (Y = 1) correctly identified. The other two cells contain the observations misclassified by the model: *False negatives* and *False positives*. The total accuracy of the predictive model is given by the sum of the *True negatives* and *True positives* cells, divided by the total number of observations. The *specificity* of the model, i.e., the ability to correctly classify negative instances, is given by the number of observations in the *True negatives* cell divided by the number of negative observations. Analogously, the *sensitivity* of the model, its ability to correctly predict positive cases, is given by the number of units in the *True positives* cells divided by the total number of positive cases.

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.techfore.2022.122016.

References

Altonji, J.G., Elder, T.E., Taber, C.R., 2005. Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. J. Polit. Econ. 113 (1), 151–184.

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., Salvestrini, V., 2018. Targeting with machine learning: an application to a tax rebate program in Italy. J. Econ. Behav. Organ. 156, 86–102.

Andini, M., Boldrini, M., Ciani, E., De Blasio, G., D'Ignazio, A., Paladini, A., 2022. Machine learning in the service of policy targeting: the case of public credit guarantees. J. Econ. Behav. Organ. 198, 434–475.

Antulov-Fantulin, N., Lagravinese, R., Resce, G., 2021. Predicting bankruptcy of local government: a machine learning approach. J. Econ. Behav. Organ. 183, 681–699.

Ash, E., Galletta, S., Giommoni, T., 2020. A machine learning approach to analyze and support anti-corruption policy. Available at. https://ssrn.com/abstract=3589545.

Athey, S., Imbens, G.W., 2017. The state of applied econometrics: causality and policy evaluation. J. Econ. Perspect. 31 (2), 3–32.

Athey, S., Imbens, G.W., 2019. Machine learning methods that economists should know about. Annu.Rev.Econ. 11, 685–725.

Ballestar, M.T., Doncel, L.M., Sainz, J., Ortigosa-Blanch, A., 2019. A novel machine learning approach for evaluation of public policies: an application in relation to the performance of university researchers. Technol. Forecast. Soc. Chang. 149, 119756.

Baysan, C., Burke, M., González, F., Hsiang, S., Miguel, E., 2019. Non-economic factors in violence: evidence from organized crime, suicides and climate in Mexico. J. Econ. Behav. Organ. 168, 434–452.

Bennett Moses, L., Chan, J., 2018. Algorithmic prediction in policing: assumptions, evaluation, and accountability. Polic. Soc. 28 (7), 806–822. https://doi.org/10.1080/10439463.2016.1253695.

Blumenstock, J., Cadamuro, G., On, R., 2015. Predicting poverty and wealth from mobile phone metadata. Science 350 (6264), 1073–1076.

Brayne, S., 2017. Big data surveillance: the case of policing. Am. Sociol. Rev. 82, 977–1008.

Brayne, S., Christin, A., 2020. Technologies of crime prediction: the reception of algorithms in policing and criminal courts. Soc. Probl. 68 (3), 608–624.

Carrieri, V., Lagravinese, R., Resce, G., 2021. Predicting vaccine hesitancy from area-level indicators: a machine learning approach. Health Econ. 30 (12), 3248–3256.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., Mullainathan, S., 2016. Productivity and selection of human capital with machine learning. Am. Econ. Rev. 106 (5), 124–127.

Chandler, D., Levitt, S.D., List, J.A., 2011. Predicting and preventing shootings among at-risk youth. Am.Econ.Rev. 101 (3), 288–292.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Chen, X., Cho, Y., Jang, S.Y., 2015. Crime prediction using Twitter sentiment and weather. In: 2015 Systems And Information Engineering Design Symposium. IEEE, pp. 63–68.

Clifton, B., Lavigne, S., Tseng, F., 2017. Predicting Financial Crime: Augmenting the Predictive Policing Arsenal. April.

De Angelis, I., de Blasio, G., Rizzica, L., 2020. Lost in corruption. Evidence from EU funding to southern Italy. <sb:contribution><sb:title>Ital.</sb:title></sb:contribution><sb:host><sb:issue><sb:series><sb:title>Econ. J.</sb:title></sb:series></sb:issue></sb:host> 1–23.

Decarolis, F., Giorgiantonio, C., 2020. Corruption red flags in public procurement: new evidence from Italian calls for tenders. In: Questioni di Economia e Finanza, Occasional Papers, 544.
Gallego, J., Rivero, G., Martínez, J., 2021. Preventing rather than punishing: an early warning model of malfeasance in public procurement. Int. J. Forecast. 37 (1), 360–377.
Garbero, A., Letta, M., 2022. Predicting household resilience with machine learning: preliminary cross-country tests. Empir. Econ. 1–14.
Garbero, A., Carneiro, B., Resce, G., 2021. Harnessing the power of machine learning analytics to understand food systems dynamics across development projects. Technol. Forecast. Soc. Chang. 172, 121012.
Gorr, W., Harries, R., 2003. Introduction to crime forecasting. Int. J. Forecast. 19 (4), 551–555.
Groff, E.R., La Vigne, N.G., 2001. Mapping an opportunity surface of residential burglary. J. Res. Crime Delinq. 38 (3), 257–278.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, And Prediction. Springer Science & Business Media.
Healy, P., Serafeim, G., 2016. In: Who Pays for White-collar Crime? Working Paper. Harvard Business School, pp. 16–148.
Horrocks, J., Menclova, A.K., 2011. The effects of weather on crime. N. Z. Econ. Pap. 45, 231–254.
Hossain, M., Mullally, C., Asadullah, M.N., 2019. Alternatives to calorie-based indicators of food security: an application of machine learning methods. Food Policy 84, 77–91.
Jansen, F., 2018. Data Driven Policing in the Context of Europe. Data Justice Lab.
Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science 353 (6301), 790–794.
Kang, J.S., Kuznetsova, P., Luca, M., Choi, Y., 2013. Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1443–1448.
Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction policy problems. Am. Econ. Rev. 105 (5), 491–495.
Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2018. Human decisions and machine predictions. Q. J. Econ. 133 (1), 237–293.
Kleinberg, J., Ludwig, J., Mullainathan, S., Sunstein, C.R., 2020. Algorithms as discrimination detectors. Proc. Natl. Acad. Sci. 117 (48), 30096–30100.
Knippenberg, E., Jensen, N., Constas, M., 2019. Quantifying household resilience with high frequency data: temporal dynamics and methodological options. World Dev. 121, 1–15.
Lantz, B., 2019. Machine Learning With R: Expert Techniques for Predictive Modeling. Packt Publishing Ltd.
Lentz, E.C., Michelson, H., Baylis, K., Zhou, Y., 2019. A data-driven approach improves food insecurity crisis prediction. World Dev. 122, 399–409.
Lima, M.S.M., Delen, D., 2020. Predicting and explaining corruption across countries: a machine learning approach. Gov. Inf. Q. 37 (1), 101407.
López-Iturriaga, F.J., Sanz, I., 2018. Predicting public corruption with neural networks: an analysis of Spanish provinces. In: Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, vol. 140(3). Springer, pp. 975–998. December.
Mastrobuoni, G., 2020. Crime is terribly revealing: information technology and police productivity. Rev. Econ. Stud. 87 (6), 2727–2753.
Matsuura, K., Willmott, C.J., 2015. Terrestrial Air Temperature And Precipitation: Monthly And Annual Time Series (1900 - 2014), vol. 4.01.
McBride, L., Nichols, A., 2018. Retooling poverty targeting using out-of-sample validation and machine learning. World Bank Econ. Rev. 32 (3), 531–550.
Meijer, A., Wessels, M., 2019. Predictive policing: review of benefits and drawbacks. Int. J. Public Adm. 42 (12), 1031–1039.
Mocetti, S., Rizzica, L., 2019. Criminalità organizzata e corruzione: incidenza e effetti sull'economia reale in Italia. In: Rassegna Economica, vol. 82, pp. 85–107.
Mohler, G., 2014. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. Int. J. Forecast. 30 (3), 491–497.
Mullainathan, S., Spiess, J., 2017. Machine learning: an applied econometric approach. J. Econ. Perspect. 31 (2), 87–106.
Oswald, M., Grace, J., Urwin, S., Barnes, G.C., 2018. Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. Inf.Commun.Technol.Law 27 (2), 223–250. https://doi.org/10.1080/13600834.2018.1458455.
Perez, A., Ganguli, S., Ermon, S., Azzari, G., Burke, M., Lobell, D., 2019. Semi-supervised multitask learning on multispectral satellite images using Wasserstein generative adversarial networks (GANs) for predicting poverty. arXiv preprint arXiv:1902.11110.
Perry, W.L., McInnis, B., Price, C.C., Smith, S.C., Hollywood, J.S., 2013. Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. RAND Corporation, Santa Monica, Washington, Pittsburgh, New Orleans, Jackson, Boston, Doha, Cambridge, Brussels.
Putnam, R.D., Leonardi, R., Nanetti, R.Y., 1992. Making democracy work: Civic traditions in modern Italy. Princeton university press.
Ranson, M., 2014. Crime, weather, and climate change. J. Environ. Econ. Manag. 67 (3), 274–302.
Rockoff, J.E., Jacob, B.A., Kane, T.J., Staiger, D.O., 2011. Can you recognize an effective teacher when you recruit one? Educ.FinancePolicy 6 (1), 43–74.
Varian, H.R., 2014. Big data: new tricks for econometrics. J. Econ. Perspect. 28 (2), 3–28.
Zhao, Y., Chen, Y., 2013. Data Mining Applications With R. Academic Press.

Guido de Blasio Guido de Blasio is an Economist at the Bank of Italy, Department of Economics and Statistics. His research mainly focuses on regional science and urban economics, program evaluation, cultural economics (social capital), applied econometrics and machine learning.

Alessio D'Ignazio Alessio D'Ignazio is an Economist at the Bank of Italy. His research mainly concerns the structural aspects of the Italian economy. He has worked on issues relating to regional economic disparities, the functioning of the credit market and policy evaluation, publishing several papers in national and international referred journals. He holds a MSc from the University of York.

Marco Letta Marco Letta is an Assistant Professor of Economics at the Department of Social Sciences and Economics, Sapienza University of Rome. His main research fields are regional economics, development economics, policy evaluation, applied econometrics and machine learning.