ISSN 2385-2755
DiSSE Working papers
[online]

SAPIENZA
UNIVERSITÀ DI ROMA

WORKING PAPERS SERIES
DIPARTIMENTO DI SCIENZE SOCIALI ED ECONOMICHE

Predicting Corruption Crimes with Machine Learning. A Study for the Italian Municipalities

Guido de Blasio, Alessio D'Ignazio, Marco Letta

N. 16/2020

SAPIENZA - UNIVERSITY OF ROME
P.le Aldo Moro n.5 – 00185 Roma T(+39) 0649910563
CF80209930587 – P.IVA 02133771002

Predicting Corruption Crimes with Machine Learning. A Study for the Italian Municipalities*

Guido de Blasio,° Alessio D'Ignazio,• Marco Letta±

Preliminary version: October 7, 2020

Abstract

Using police archives, we apply machine learning algorithms to predict corruption crimes in Italian municipalities during the period 2012-2014. We correctly identify over 70% (slightly less than 80%) of the municipalities that will experience corruption episodes (an increase in corruption crimes). We show that algorithmic predictions could strengthen the ability of the 2012 Italy's anti-corruption law to fight white-collar delinquencies.

Keywords: crime prediction, white-collar crimes, machine learning, classification trees, policy targeting

JEL Codes: C52, D73, H70, K10

* We are grateful to Fabrizio Balassone, Augusto Cerqua, Giovanni Cerulli, Arthur Charpentier, Gianluca Maria Esposito, Emmanuel Flachaire, Sauro Mocetti, Pierluigi Montalbano, Alessio Muscarnera and Lucia Rizzica for helpful comments and suggestions. We also thank seminar participants to Sapienza University of Rome and the 2020 AISRe Conference. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Institutions they are affiliated with.
° Bank of Italy; guido.deblasio@bancaditalia.it
• Bank of Italy; alessio.dignazio@bancaditalia.it
± Sapienza University of Rome; marco.letta@uniroma1.it *(Corresponding Author)*

"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, June 4 2016

1. Introduction

Corruption in Italy is a major problem. According to the index of Transparency International,[1] which inversely ranks 180 countries by their perceived levels of public sector corruption as stated by experts and businesspeople, in 2018 Italy was in the 52nd position, delayed from Germany (11th), France (21st), and Spain (41st). Corruption is unevenly distributed across the country: 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), having a map of the areas where to concentrate investigation efforts should be considered a priority.

This paper uses machine learning (ML) algorithms in an attempt to provide such a map. 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). Applications of ML algorithms 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 over judges' decision 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);

[1] See https://www.transparency.org/en/.

improve poverty (Blumenstock et al., 2015; Jean et al., 2016; McBride & Nichols, 2018; Perez et al., 2019) and food insecurity (Hossain et al., 2019; Knippenberg et al., 2019; Lentz et al., 2018) targeting; enhance the effectiveness of public programs (Andini et al., 2018; Andini et al., 2019). More specifically linked to the content of this paper, another active area of research and application of ML can be found in criminology and goes under the heading of 'predictive policing'. The idea is that of predicting (and preventing) crime before it happens, in order to reduce criminality and use public resources more efficiently.

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 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 prediction 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) and 479-481 (crimes against public faith) 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. 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 predict from 70% to 80% of the local corruption. The prediction depends on the values of a few variables, primarily referring to the characteristics of the local labour and housing markets, and the previous history of white-collar crimes. We also discuss how ML predictions can be used to strengthen the effectiveness of the 2012 Italy's anti-corruption law. For instance, the law envisages a stricter anti-corruption regulation for 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. We also discuss a number of issues related to the adoption of ML algorithms to fight corruption. 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. We also underscore that ML adoption could provide higher policy standards in terms of both transparency and bias reduction.

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

The idea of being able to predict (and prevent) crime before it happens is gaining increasing interest across both researchers and police forces (Brayne & Christin, 2020; Meijer and Wessels, 2019). To this aim, several techniques have been used and refined over time (see Grover et al., 2007). A first set of methods involves purely statistical approaches to predict crime by means of individual "profiling". Such methods rely on individual characteristics, such as their social connections and other personal data on behaviours. A second set of methods emphasizes the spatial clustering of criminal activity and leads to the identification of the so-called hot spots, i.e., areas where offenders tend to repeat their crime. Both approaches share some drawbacks, as they both rely on a time-consistency assumption: in the first case, the

unchanging modus operandi of the offender; in the second, the persistence of the same area as a crime target. As argued by Brayne (2017), the availability of big data and ML techniques has recently allowed to moving from hotspot policing to predictive policing, i.e., the identification of police targets to prevent or solve past crimes through statistical predictions (Perry et al., 2013). Previous work on predictive policing mainly refers to the US and crimes such as burglaries, thefts, violence against the person (see Meijer and Wessels, 2019, and Bennett and Chan, 2018, for a review). Mohler et al. (2015) employ an ML-based model to predict where crimes will take place and test the effectiveness of such predictions against current hotspot mapping practice through a random experiment involving two divisions of Los Angeles (US) and Kent (UK) police. They show that their models predict 1.4–2.2 times as much crime compared to a dedicated crime analyst using existing criminal intelligence and hotspot mapping practice. Further, there are some recent applications focusing on European countries too: Mastrobuoni (2020), for instance, employs a 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.

Machine learning tools are now widely implemented in predictive policing across the US. A notable example is Chicago where, in 2013, an algorithm was released to predict who is more likely to be involved in shooting (Strategic subject list) in order to prioritize resources to focus on individuals at highest risk. In Europe, the use of algorithms in policing is at an embryonal level and mostly involves the United Kingdom. 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). Finally, another recent work (Wheeler & Steenbeek, 2020) employs the random forest method to produce long-term crime forecasts for robberies in Dallas.

Differently from other types of crimes, white-collar offences have been scantily studied. López-Iturriaga and Sanz (2018) refer to the case of Spanish provinces and use info on corruption episodes reported by the media or that went to court between 2000 and 2012 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 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 machine and artificial neural networks, on cross-country data

to identify the most important predictors of corruption at the country level. Gallego et al. (2020) 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 novel data concerning the procurement of public works to predict indicators of corruption risk, showing the potential of the flexibility of such algorithms in detecting corruption in public procurement.

ML tools aiming at preventing white-collar crimes are little used. In Mexico, in order to tackle corruption in public procurement, a corruption risk index was devised for about 1500 buying units. Moreover, the Tax Administration Service has tested the effectiveness of machine learning algorithms to 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[2]. In Ukraine, an ML tool was recently launched to fight corruption and fraud in public procurement. Such a tool identifies tenders with a high risk of corruption; when tenders or purchases are flagged by the tool, they are reported to the authorities to be investigated.

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 (8,049 out of 8,092 municipalities); data on white-collar crimes, which is the main object of our study, is available for 7,794 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. We want to 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 1,000 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

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

year, and 0 otherwise.[3] Figures 1 and 2 provide maps for the value of our two target variables in the year 2012.

[Figure 1]

[Figure 2]

The set of predictors consists of socio-economic, demographic, geographic and biophysical characteristics, available at the municipality level. More specifically, we employ three different data sources: 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;[4] 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 regions (namely Southern Europe, Eastern Europe and Northern Africa); data on the number of police stations within 50 kilometres from the municipality centroid and the Euclidean distance from the centroid to the closest police station;[5] climate data from the University of Delaware weather database (Matsuura & Willmott, 2015),[6] to control for biophysical and climatic heterogeneity across the country.[7] As a result, we are able to assemble a dataset with approximately 100 features or predictors. We also employ as additional predictors lagged (2008-2011) values of the outcome variables. Table 1 reports summary statistics for a selected list of features. The average municipality population is roughly equal to 7,500 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. Therefore, when we employ our machine learning algorithm to inform the actual anti-corruption strategy, we will be using only features available to the Italian policymaker at the time of the ratification of the law, i.e., in 2012.

[3] 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.
[4] The 8 Mila Census database is publicly available at the following link: http://ottomilacensus.istat.it/.
[5] Freely available here.
[6] The raw data on police stations at the local level are available here.
[7] 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 & Menclova (2011), Ranson (2014), Chen et al. (2015) and Baysan et al. (2018).

[Table 1]

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.

[Table 2]

3.2 Methods

ML techniques use highly flexible functional forms. The degree of flexibility is the result of a well-known 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 the standard ML routine, 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). In order to solve the bias-variance trade-off (Hastie et al., 2009), we also 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 2011 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 ML algorithm is the classification tree (Hastie et al., 2009). Classification trees 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 very appealing for policy targeting purposes. From a technical point of view, the algorithm divides the data into progressively smaller subsets to identify patterns that

can be used for predicting a specific binary 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 that reduces the risk of overfitting. In fact, a high number of levels in a tree is likely to overfit the data, leading to a predictive model which performs very well in-sample, but poorly out-of-sample. 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. As specified above, we select the optimal value of *cp* via 10-fold cross-validation in the training dataset, which in our case corresponds to the complexity parameter value of the unpruned tree.

Before performing our classification exercise, we need to tackle the challenge stemming our highly imbalanced dataset. Our outcome variables are both highly skewed toward zeros (see Figures 1 and 2 and Table 2). In the case of imbalanced datasets, ML 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 their prediction performance is dominated by the accurateness of predicting the over-represented label (y=0 in our case). Using the machine learning jargon, predictive exercises on imbalanced datasets result in a very high *specificity* (i.e., true negatives) but an extremely low, if not null, *sensitivity* (true positives).

This is visible in our case by looking at the prediction accuracy we obtain by using the original sample: Tables A.1 and A.2 in the Appendix 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. This technique oversamples the under-represented cases *and* undersamples the majority class, leading to a smaller rebalanced dataset. To oversample to minority class, SMOTE generates new synthetic observations by considering the *k* (10, In our case) nearest neighbours of each minority class sample (Chawla et al., 2002). 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 prediction is tested on the original skewed sample. After rebalancing our training dataset, the two outcomes are almost perfectly

balanced between the two classes, and the sample size is sharply reduced due to the undersampling of the majority class.[8] On these rebalanced data, we then perform our decision tree algorithm, whose results are provided in the next section.

4. Results

Figure 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 7,390 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%. Figure 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 7,361 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%.

[Figure 3]

[Figure 4]

Tables 3 and 4 highlight the prediction accuracy of the classification trees described in Figures 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. 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.1 (where sensitivity for the 2012 sample is 31.2%, while the overall accuracy is

[8]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).

92.6%).[9] 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.2, 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.[10]

[Table 3]

[Table 4]

ML algorithms use highly non-linear functional forms. However, endowed with the same set of predictors used to produce the classifications trees of Figures 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.3 and A.4 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.[11] 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.5. We find only a limited reduction in overall accuracy. Unexpectedly, we observe that, when we do not consider previous corruption episodes, the sensitivity of the ML predictions increases.

5. Algorithms in the service of the anti-corruption law

In 2012 the law 190,[12] named "Rules for the prevention and repression of corruption and unlawfulness in public administration" and informally known as "Legge Severino" (from the

[9] 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.
[10] There is no corresponding tree for this table on the pre-SMOTE classification performance because there is no tree (the algorithm predicts all zeroes).
[11] 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.
[12] See http://www.anticorruzione.it/portal/public/classic/MenuServizio/FAQ/Anticorruzione.

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.*[13] The rationale behind the decision was that of excluding smaller municipalities where the costs related to the regulation were larger than the associated benefits because those municipalities would have experienced very few cases of white-crime episodes anyway. Smaller municipalities also receive fewer public resources, and this makes them, in principle, less exposed to corruption risk. According to new estimates provided by De Angelis et al. (2020), the impact of law 190 seems to be favourable, at the least in the South of Italy, where the municipalities over 15 thousand residents experienced fewer corruption episodes linked to the EU regional transfers.

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.[14] 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. 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 prediction, 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%.

[Table 5]

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

[14] 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]

6. 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. We tried with random forest, but we did not get significant improvements over the accuracy of the classification trees.[15] However, we did not try more complex ML algorithms, such as support vector machine or neural networks. Another important aspect refers to the number of features that we have used to train the algorithms. We have a long list of variables, but still quite restricted when compared with the amount of information (Big Data) that in other ML applications have been exploited. Indeed, the relatively low number of predictors and observations might be the main reason why the performance of the classification tree is competitive with that of more sophisticated techniques such as the random forest. With more features, we could have obtained an even higher accuracy. One aspect is that we have only a limited subset of the information available to police departments. For instance, having a longer time series could have allowed exploiting additional features of time-varying nature. Additionally, more detailed georeferenced data could permit intra-municipality predictions. Note that the availability of information at the local level is going to increase over time: future predictions are going to be more precise than the ones we have provided in this paper.

When compared with the law 190/2012, we have been able to improve predictive sensitivity (y=1) at the cost of losing accuracy in specificity (y=0). Given the high socio-economic costs of corruption, the benefits related to higher sensitivity seem to be warranted. However, the pool of municipality 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, then the overall regulatory costs would, in principle, rise. However, in the current circumstances, it is difficult to imagine that a legislative act could delegate the identification of municipalities to an algorithm. More realistically, ML predictions could be used to refine the existing regulation, ruling out municipalities above the 15,000 inhabitants that are not predicted to be corruption-

[15] Results are available upon request.

prone. Under these circumstances, 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.

One aspect on which the literature on ML has concentrated a lot refers to bias. Suppose that our data are contaminated because corruption episodes are more likely to be reported in certain communities than others. Think, for instance, of a social capital story (Putnam, 1993). 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 y=1, *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 post-SMOTE sample (the one on which the prediction is based) is likely to be less exposed to contamination, compared to the original sample. What SMOTE does is an undersampling of the most numerous class, in our case the zeros, by excluding those observations which are less similar – as measured by comparing observable features - to the other class, the ones. 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.

We have picked up classification trees purposely to minimize transparency issues. The trees described in Figures 3 and 4 can be easily communicated to the public. Obviously, it is easier to understand one single threshold than a bunch of them, sequentially linked. However, this cost might be considered not that large especially because the prediction based on the trees increases effectiveness in finding out 'corrupted' municipalities and thus communicated as necessary to serve a public aim (in this respect, the population threshold implied by the law

does not have such a sound foundation). 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 (Figure 3) or 6 (Figure 4) variables. This is not the case for more complex algorithms, which require the whole information set at any step. Finally, and still related to transparency, ML methods highlight the targeting that an authority interested in fighting corruption should adopt. Therefore, they can also provide information on whether other 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 have police investigations far from some places. Having the ML prediction map, which can be easily compared with that of the actual police efforts, might shed light on that.

In conclusion, our findings suggest that the combination of new data and data-driven machine learning techniques might provide innovative and impartial tools to help the policymaker improve ex-ante targeting and regulatory design, an exceptionally delicate and critical task in the fight to white-collar delinquencies.

References

Altonji, J. G., & Elder, T. E., & Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, *113*(1), 151-184.

Andini, M., Boldrini, M., Ciani, E., de Blasio, G., D'Ignazio, A., & Paladini, A. (2019). Machine learning in the service of policy targeting: the case of public credit guarantees. *Bank of Italy Temi di Discussione (Working Paper) No*, *1206*.

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. *Journal of Economic Behavior & Organization*, *156*, 86-102.

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. *Journal of Economic Perspectives*, *31*(2), 3-32.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685-725.

Baysan, C., & Burke, M., & González, F., & Hsiang, S., & Miguel, E. (2018). Economic and non-economic factors in violence: Evidence from organized crime, suicides and climate in Mexico. *National Bureau of Economic Research Working Paper* No. 24897.

Bennett Moses, L., & Chan, J. (2018). Algorithmic prediction in policing: assumptions, evaluation, and accountability, Policing and Society, 28:7, 806-822, DOI: 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. American Sociological Review, 82, 977–1008.

Brayne, S., & Christin, A. (2020). Technologies of Crime Prediction: The Reception of Algorithms

in Policing and Criminal Courts. *Social Problems,* spaa004.

Chalfin, A., & Danieli, O., & Hillis, A., & Jelveh, Z., & Luca, M., & Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124-27.

Chandler, D., & (Levitt, S. D., & List, J. A. (2011). Predicting and preventing shootings among at-risk youth. *American Economic Review*, *101*(3), 288-92.

Chawla, N. V., & Bowyer, K. W., & Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *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* (pp. 63-68). IEEE.

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

De Angelis, I., & de Blasio, G., & Rizzica, L. (2020). Lost in Corruption. Evidence from EU Funding to Southern Italy. *Italian Economic Journal*, 1-23.

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

Gallego, J., Rivero, G., & Martínez, J. (2020). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting*.

Grover V., & Adderley R., & Bramer M. (2007). Review of Current Crime Prediction Techniques. In: Ellis R., Allen T., Tuson A. (eds) Applications and Innovations in Intelligent Systems XIV. SGAI 2006. Springer, London.

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). Who Pays for White-Collar Crime? Working Paper 16-148, Harvard Business School.

Horrocks, J., & Menclova, A.K. (2011). The effects of weather on crime. *New Zealand Economic*

Papers, Vol, 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.

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. *Proceedings of the 2013 conference on empirical methods in natural language processing,* 1443-1448.

Kleinberg, J., & Lakkaraju, H., & Leskovec, J., & Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237-293.

Kleinberg, J., & Ludwig, J., & Mullainathan, S., & Obermeyer, Z. 2015. Prediction policy problems. *American Economic Review*, *105*(5), 491-95.

Knippenberg, E., & Jensen, N., & Constas, M. (2019). Quantifying household resilience with high frequency data: Temporal dynamics and methodological options. *World Development*, 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 Development*, 122, 399-409.

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

López-Iturriaga, F. J., & Sanz, I., (2018). Predicting Public Corruption with Neural Networks: An Analysis of Spanish Provinces, Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, Springer, vol. 140(3), pages 975-998, December.

Mastrobuoni, G. (2020). Crime is terribly revealing: Information technology and police productivity. *The Review of Economic Studies*, rdaa009.

Matsuura, K., & Willmott, C..J. (2015). Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1900 - 2014), v 4.01.

McBride, L., & Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. The World Bank Economic Review, 32(3), 531-550.

Meijer, A., & Wessels, M. (2019). Predictive Policing: Review of Benefits and Drawbacks, International Journal of Public Administration, 42:12, 1031-1039, DOI: 10.1080/01900692.2019.1575664

Mocetti, S., & Rizzica, L. (2019**).** Criminalità organizzata e corruzione: incidenza e effetti sull'economia reale in Italia**,** *Rassegna Economica***,** vol. 82, 85-107.

Mohler, G. O., & Short, M. B., & Malinowski, S., & Johnson, M., & Tita, G. E., & Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing. *Journal of the American statistical association*, *110*(512), 1399-1411.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *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, Information & Communications Technology law, 27:2, 223-250, DOI: 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. Santa Monica, Washington, Pittsburgh, New Orleans, Jackson, Boston, Doha, Cambridge, Brussels: RAND Corporation.

Ranson, M. (2014). Crime, weather, and climate change. *Journal of environmental economics and management*, 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?. *Education finance and Policy*, *6*(1), 43-74.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28.

Wheeler, A. P., & Steenbeek, W. (2020). Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology*, 1-36.

Table 1

Descriptive statistics – Selected features

Variable	Year	Mean	Var	sd	Count
Population	2011	7472.740	1.637e+09	40457.589	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	19856.800	140.914	7071
Vulnerability index	2011	98.770	2.701	1.644	7794
Place in vulnerability index ranking	2011	4036.333	5454224.210	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	96898.239	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 1,000 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 1,000 inhabitants) has increased with respect to the previous year, and 0 otherwise.

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 1,000 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 1,000 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 1,000 inhabitants) is positive and 0 otherwise.

Table 6

Anti-corruption threshold *vs* decision tree rule performances

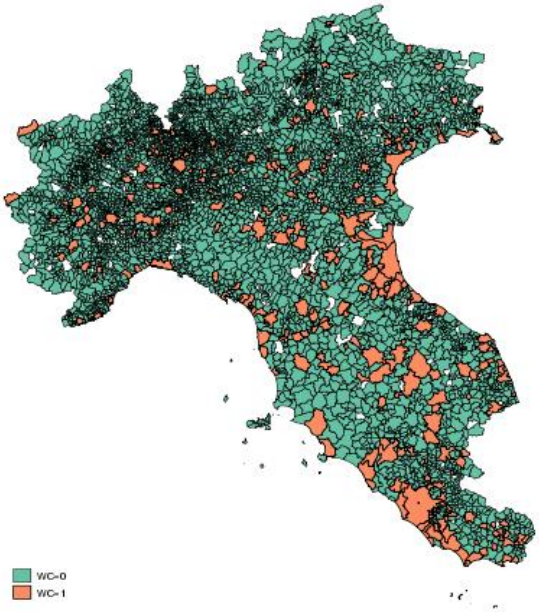
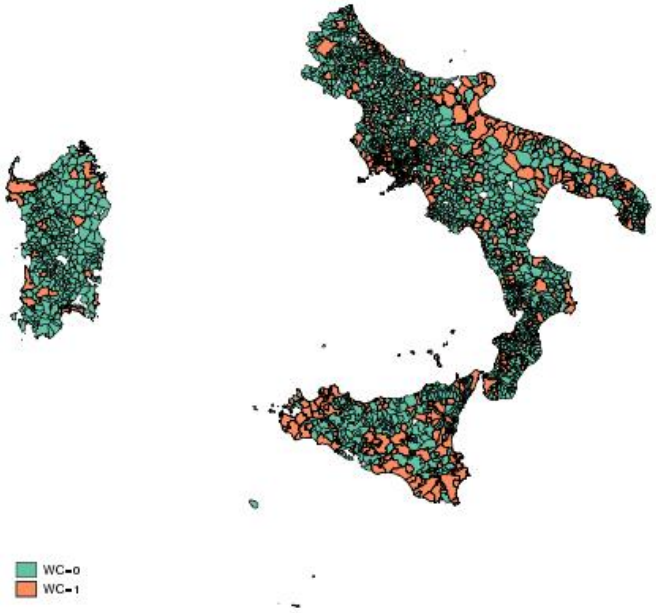
(testing sample; variable: Δ WC crime rate)

		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 1,000 inhabitants) has increased with respect to the previous year, and 0 otherwise.

Figure 1

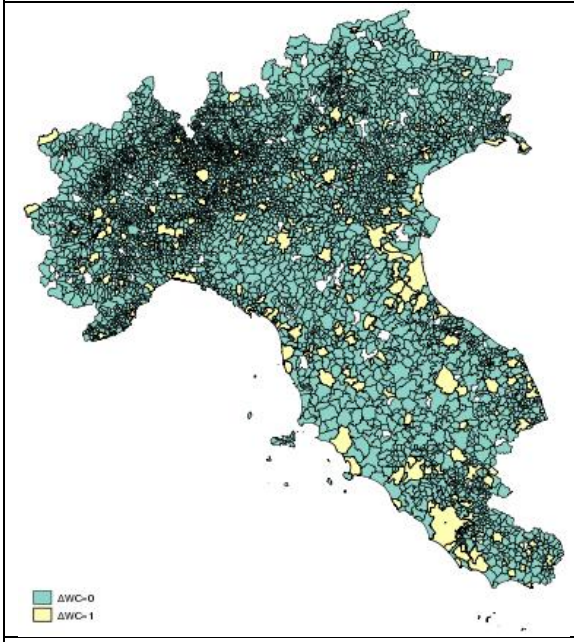
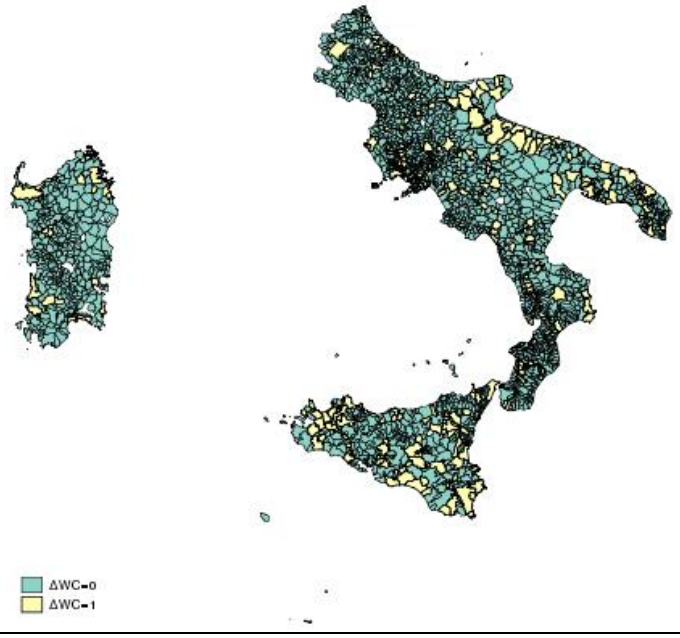
WC crime rate across Italian municipalities – 2012

Centre and North	South and Islands
	

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.

Figure 2

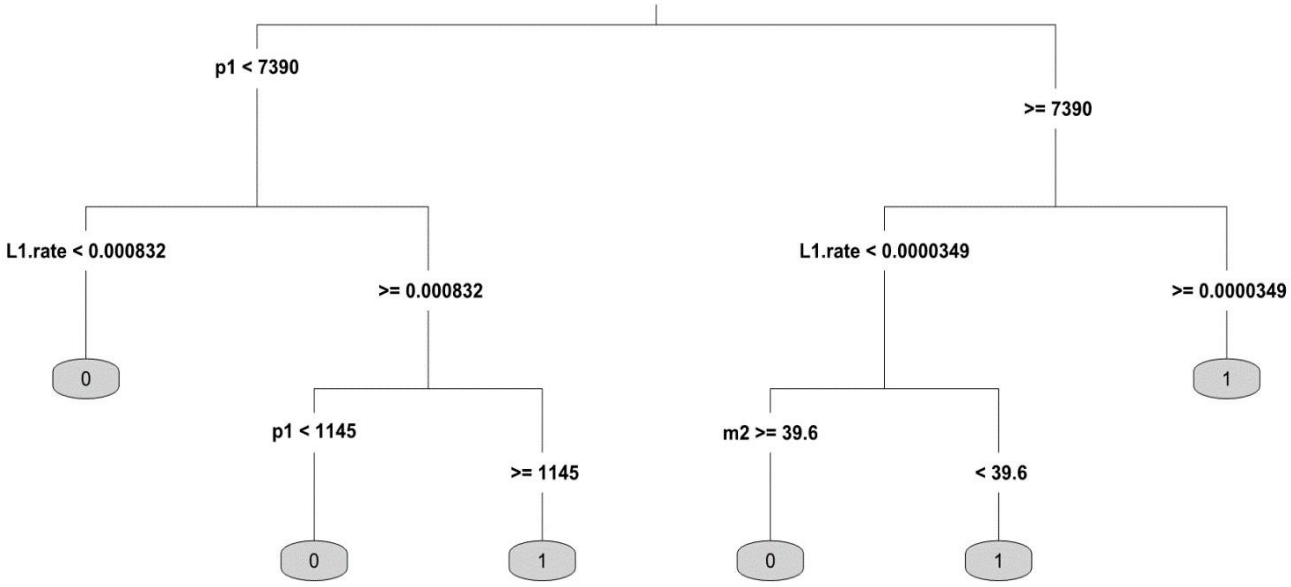
ΔWC crime rate across Italian municipalities – 2012

Centre and North	South and Islands
	

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.

Figure 3

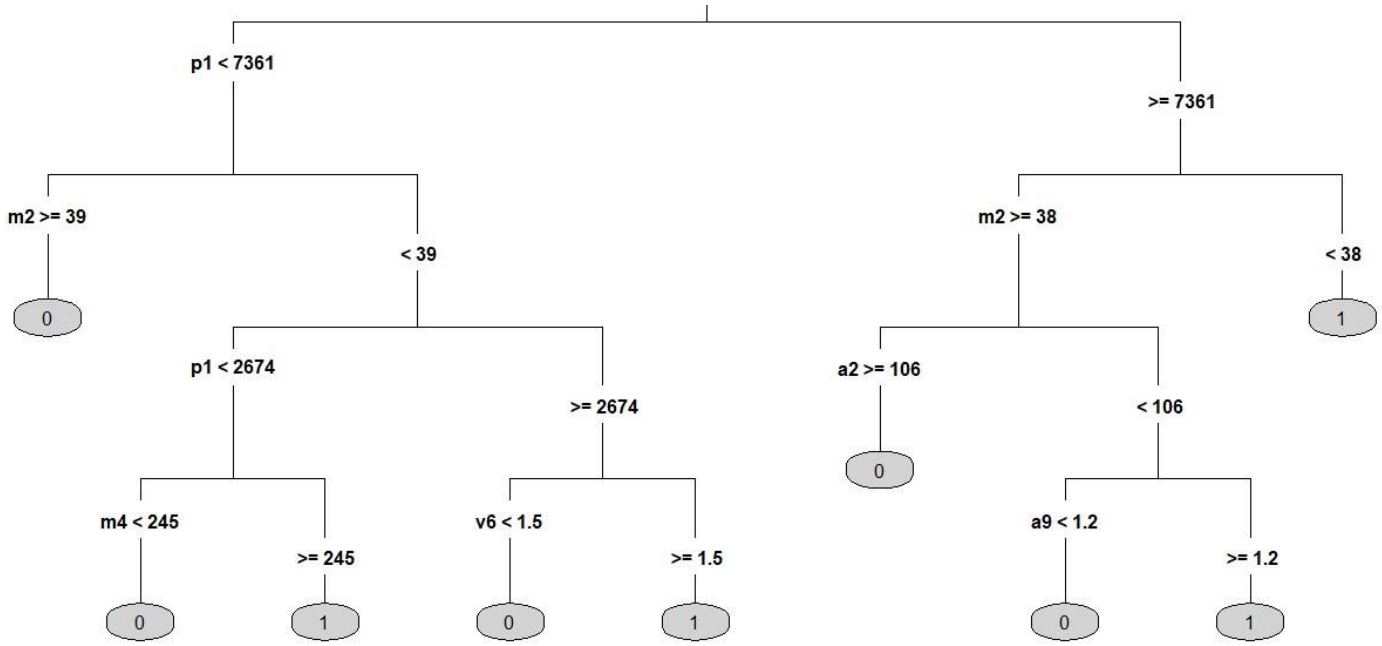
Classification tree for WC Crime Rate – Post-SMOTE data



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)

Figure 4

Classification tree for Δ WC Crime Rate – Post-SMOTE data



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 (%)

Appendix

Table A.1

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 1,000 inhabitants) is positive and 0 otherwise.

Table A.2

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 1,000 inhabitants) has increased with respect to the previous year, and 0 otherwise.

Table A.3

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 1,000 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.4

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 1,000 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*.

Table A.5

Post-SMOTE Decision tree performance on the testing sample

***without* lagged crime predictors**

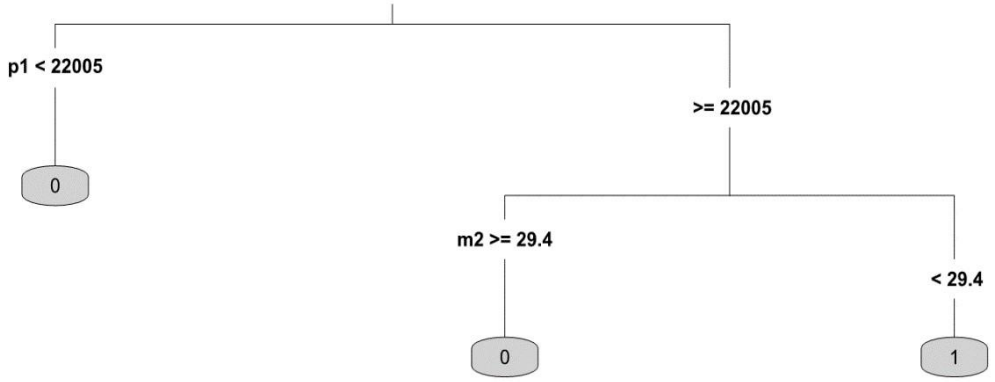
(variable: WC crime rate)

		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 1,000 inhabitants) is positive and 0 otherwise.

Figure A.1

Classification tree for WC crime rate – Pre-SMOTE data



Legend	
p1:	Population
m2	Daily mobility outside the municipality for study or work (share of the working-age population)