BANCA D'ITALIA
EUROSISTEMA

UIF
Unità di Informazione Finanziaria per l'Italia

Quaderni dell'antiriciclaggio

Analisi e studi

A machine learning approach for the detection of firms linked to organised crime in Italy, based on balance sheet data

Pasquale Cariello, Marco De Simoni, Stefano Iezzi

giugno 2024

numero 22

BANCA D'ITALIA
EUROSISTEMA

Unità di Informazione Finanziaria per l'Italia

Quaderni dell'antiriciclaggio

Analisi e studi

A machine learning approach for the detection of firms linked to organised crime in Italy, based on balance sheet data

Pasquale Cariello, Marco De Simoni, Stefano Iezzi

Numero 22 - Giugno 2024

La serie Quaderni dell'antiriciclaggio *ha la finalità di presentare dati statistici, studi e documentazione su aspetti rilevanti per i compiti istituzionali della UIF — Unità d'Informazione Finanziaria per l'Italia, Banca d'Italia.*

La serie si articola in due collane: la collana Dati statistici *presenta, con periodicità semestrale, statistiche sulle segnalazioni ricevute e informazioni sintetiche sull'operatività dell'Unità; la collana* Analisi e studi *comprende contributi sulle tematiche e sui metodi in materia di contrasto al riciclaggio e al finanziamento del terrorismo.*

La collana Analisi e studi *comprende lavori realizzati all'interno della UIF, talvolta in collaborazione con altri settori della Banca d'Italia o con Istituzioni esterne. I lavori pubblicati riflettono esclusivamente le opinioni degli autori, senza impegnare la responsabilità delle Istituzioni di appartenenza.*

**Unità di Informazione Finanziaria
per l'Italia**

Direttore responsabile
Enzo Serata

Per la pubblicazione cartacea:
autorizzazione del Tribunale di Roma n. 1942013 del 30 luglio 2013
Per la pubblicazione telematica:
autorizzazione del Tribunale di Roma n. 1932013 del 30 luglio 2013

ISSN 2283-3498 (stampa)
ISSN 2283-6977 (online)

A MACHINE LEARNING APPROACH FOR THE DETECTION OF FIRMS LINKED TO ORGANISED CRIME IN ITALY, BASED ON BALANCE SHEET DATA

by Pasquale Cariello, Marco De Simoni and Stefano Iezzi*

Abstract

We develop a machine learning algorithm designed to detect firms that may have connections with organized crime (OC). To this end, we utilize a firm-level dataset for Italy, merging financial information from various sources, mainly public balance sheets. We compare a sample of over 28,000 Italian firms that are highly likely to be linked to OC with randomly selected samples of allegedly lawful firms to train and test the model. Based on out-of-sample test set, the algorithm successfully identifies approximately 76% of the OC-linked firms (recall) and 74% of the allegedly lawful firms (specificity). The primary output of the algorithm is a risk score, which might be applied at an operational level (for example, as a preliminary screening tool) for supporting the action of anti-money laundering authorities and law enforcement agencies.

Sommario

In questo studio viene sviluppato un algoritmo di *machine learning* per rilevare aziende potenzialmente collegate alla criminalità organizzata (CO). A questo scopo, si utilizza un dataset di imprese italiane ottenuto integrando informazioni finanziarie provenienti da varie fonti, tra cui dati di bilancio. Per addestrare e testare il modello un campione di oltre 28.000 aziende italiane caratterizzate da una elevata probabilità di essere collegate alla CO viene confrontato con sottoinsiemi di aziende presumibilmente "sane" selezionati casualmente. I risultati ottenuti mostrano che, in fase di test, l'algoritmo identifica con successo circa il 76% delle aziende collegate alla CO (*recall*) e il 74% delle aziende presumibilmente "sane" (*specificity*). Il principale *output* dell'algoritmo è un punteggio di rischio, che potrebbe essere utilizzato a livello operativo per supportare l'azione delle autorità anti-riciclaggio e delle forze dell'ordine (ad esempio, come strumento di *screening* preliminare).

JEL Classification: C49, D22, E26, K42.
Keywords: Organized crime, infiltration, money laundering, economic crime, financial statements, machine learning.

* Bank of Italy, Financial Intelligence Unit.

Contents

1. *Introduction*[1]

The financial power of criminal organizations poses a significant threat to the economies of countries worldwide. According to estimates from the United Nations Office on Drugs and Crime, in 2009 the revenues generated by organized crime (OC) accounted for 3.6% of the world's GDP (UNODC, 2011). A study conducted by the European Commission (2021) reveals that the combined annual revenues of the nine primary criminal markets[2] in the EU ranged from €92 billion to €188 billion in 2019.

Taking a closer look at Italy, a research conducted by Transcrime in collaboration with the Italian Ministry of the Interior in 2015 indicates that the proceeds from mafia-related activities could potentially amount to as much as 2% of the nation's GDP (Transcrime, 2015).

Within this context, a paramount concern for both national and international authorities centres around the ever increasing investment of OC in the official economy and the escalating integration of organized crime groups into the legitimate economy by infiltrating and influencing lawful businesses.

Infiltrated businesses refer to entities that, although formally registered and appearing to engage in legitimate operations, are under the control of criminal organizations.[3] In infiltrated firms, there is a deep intertwining of legal and illegal activities, with legal operations primarily serving as a means to legitimize and amplify the profits generated from illicit activities.

However, organized crime achieves its goals not only by controlling infiltrated businesses but also by cultivating complex ties with legitimate enterprises. Law-abiding entrepreneurs increasingly engage willingly in transactions with criminal syndicates, driven

[1] The views expressed in this paper are those of the authors and do not necessarily represent those of the Financial Intelligence Unit for Italy, or those of Banca d'Italia. We are grateful to Mario Gara, Domenico Marchetti, Claudio Pauselli, Daniele Piazza, seminar participants at Banca d'Italia and UIF, the participants to the Conference "Central bankers go data driven" organized by the De Nederlandsche Bank, and many other colleagues for their very useful comments. Special thanks to Giuseppe Ragusa (Sapienza) for his valuable suggestions. We would also like to thank the Special Currency Police Unit (NSPV) of *Guardia di Finanza* for a collaboration on validating a preliminary version of the model, and the Italian Agency for the Administration and Destination of Seized and Confiscated Assets (ANBSC) for help on their data.

[2] Illicit drugs, trafficking in human beings, smuggling of migrants, fraud, environmental crime, illicit firearms, illicit tobacco, cybercrime and organised property crime.

[3] In a strict definition of infiltration, three distinct factors differentiate such firms from non-infiltrated ones (Ravenda et al., 2015; De Simoni, 2022): i) individuals associated with criminal organizations own the company or hold key management roles; ii) the firm's financial resources are derived from illicit activities either in part or entirely; iii) the firm's business practices often involve violence, intimidation, corruption, and other criminal conduct.

by a mutual pursuit of enhanced profits. This shift from overt coercion to subtler manipulation highlights the adaptability of organized crime in exploiting economic vulnerabilities and fostering relationships with seemingly legitimate entities, challenging traditional law enforcement strategies (*Direzione Investigativa Antimafia*, 2022).

Thus, the wider set of OC-linked firms includes not only firms directly controlled by organized crime, but also firms — owned or managed by persons *external* to OC — which 'simply' collude with organized crime, finding it profitable (needless to say, the boundary between the two categories is not always clear). Our training sample presumably includes not only strictly infiltrated firms, abut also colluding ones (see below, Section 3). For the sake of simplicity, throughout the entire paper we will use the terms OC-linked and infiltrated interchangeably, in order to identify the wider set of OC-linked firms.

Recent academic research has been dedicated to understanding how infiltration impacts a firm's financial records and to identifying the management characteristics of OC-linked firms as opposed to lawful ones. These studies consistently reveal that infiltrated firms exhibit distinct features in their financial statements. Insights from this body of literature have given rise to the development of several statistical models capable of distinguishing between infiltrated and non-infiltrated firms on the basis of financial reporting data.

The objective of this paper is to introduce a novel analytical approach for the identification of OC-linked firms. Our work innovates from three different perspectives.

We employ a unique sample of OC-linked firms built by combining both public and confidential sources. This approach distinguishes our research from previous studies in this field, as most studies have identified infiltrated firms through educated guesses based on conjectures that are challenging to verify empirically. Based on our publicly available and confidential sources, we have identified over 28,000 firms having a high likelihood of infiltration.

We compile a distinctive dataset of Italian firms (precisely, all corporations registered in Italy) spanning from 2010 to 2021. This dataset is the result of merging various sources, including financial statement data from the National Official Business Register, confidential information on firm indebtedness with the banking and financial system from the Central Credit Register of the Bank of Italy, employment data provided by the National Institute of Social Security, and details regarding owners, directors, and other firm characteristics sourced from the Chambers of Commerce. The breadth and diversity of these data sources provide us with an extensive array of financial variables and firm-level indicators, forming the foundation of our analysis.

We use this comprehensive dataset to develop a machine learning classifier capable of identifying legally registered businesses susceptible to OC influence. Our approach employs XGBoost (eXtreme Gradient Boosting), a widely adopted technique in both scientific and industrial domains. To the best of our knowledge, there is no prior literature exploring the application of this specific algorithm for identifying legally constituted entities that may be connected to organized crime.

The main outcome of our classification algorithm is a risk score for all capital companies active in Italy between 2010 and 2021. The aggregate results may shed light on the areas and sectors at higher risk of infiltration in Italy. The risk score can be used also at an operational level for supporting (for example, as a preliminary screening device) the action of anti-money laundering authorities and law enforcement agencies.[4]

The paper is organized as follows: Section 2 briefly presents the main findings from the literature regarding the role of OC in the legitimate economy and outlines the motivation underlying this study. Sections 3 and 4 explain how the sample of OC-linked firms is built and discuss the data, respectively. Section 5 describes the classification methodology approach and the main results, while Section 6 provides a robustness analysis. In Section 7 we discuss potential applications for AML and law enforcement agencies' fight against criminal infiltration of legal economy, while Section 8 presents the results of an external validation. Section 9 provides some concluding remarks.

2. Literature review

Numerous scholars have attempted to estimate how OC presence negatively affects the economy, for example, by hindering competition and the optimal allocation of resources, which, in turn, may reduce overall output (Peri, 2004; Barone and Mocetti, 2014; Pinotti, 2015). The literature has also examined other costs associated with OC presence, such as eroded quality of the political class (Daniele and Geys, 2015), reduced electoral competition (De Feo and De Luca, 2013), and diminished foreign investments (Daniele and Marani, 2011).

[4] The algorithm presented in this paper is the result of the evolution of an initial model based on a purely statistical approach using the technique of propensity score matching and a restricted sample of approximately 200 infiltrated companies. The model's outcomes were validated through cross-referencing with UIF suspicious transaction reports (STRs) data and a collaboration with the Special Currency Police Unit (NSPV) of the *Guardia di Finanza*, showcasing encouraging results.

The question of how infiltrated firms operate in the economy, with a particular focus on Italy, is a subject of extensive debate in the literature. Indeed, several scholars have recently engaged in explaining the effects of infiltration on Italian firms' financial statements. By examining a list of businesses subject to mafia-related legal proceedings and located in central and northern Italy, Fabrizi et al. (2017a) show that criminal companies are larger, more indebted and hold more liquid assets than legal ones. Bianchi et al. (2020) analyse companies based in Lombardy with connections to organized crime and illustrate how criminal organizations 'cannibalize' profits and deplete resources, often through money laundering schemes. Mirenda et al. (2022) investigate the infiltration of 'Ndrangheta, a criminal organization based in the Southern region of Calabria, into firms situated in the Central and Northern regions of Italy. They show that 'Ndrangheta tends to enter firms in economic and financial distress and those mostly relying on public sector procurement, ultimately resulting in higher revenues. De Simoni (2022) concludes that infiltrated firms, despite having higher revenues, are less profitable and maintain more cash assets. He also argues that investment decisions and funding strategies vary depending on the type and purpose of infiltration.

The literature on the analysis of infiltrated firms' financial statement is sufficiently wide to provide a sound enough support to our idea. Our work mainly capitalizes the findings of recent studies in order to build a highly diversified set of financial variables and indicators so as to train the machine learning algorithm to identify businesses which are possibly infiltrated. A similar methodology found in existing literature is presented by Ravenda et al. (2015), where the authors employ a logistic regression model to identify registered firms in Italy associated to mafia, on the basis of distinctive characteristics derived from their financial statements. Furthermore, there are other contributions that offer machine learning applications in the broader field of financial fraud, although these are loosely related to the focus of our study (Chengwei et al., 2015; Maka et al., 2020; Sadgali et al., 2019; Sharma and Panigrahi, 2013; Wyrobek, 2020).

3. The list of firms connected to OC

Italian civil law mandates that all corporate entities, encompassing both limited liability and joint stock companies, are mandated to publish financial statements on a yearly basis. These statements adhere to a standardized format applicable to all businesses, except for firms below a certain size threshold. These smaller businesses are granted the option to consolidate specific variables, particularly pertaining to credit and debt items.[5] Consequently, our emphasis is on private limited liability and joint stock companies. Therefore, when using the term "firms" in the paper, we specifically refer to these types of companies.

Training a classifier requires selecting a subset of firms identified as infiltrated. The categorization of a firm as infiltrated, along with the timeframe during which this status applies, is derived from an array of distinct sources.

The first subset, referred to as 'list A,' includes 229 firms and has been defined in collaboration with a specialized Italian law enforcement unit focused on combating organized crime and terrorism (De Simoni, 2022). This selection of firms encompasses businesses seized or confiscated by judicial bodies as a result of the main anti-mafia investigations conducted in Italy in the period 2007 to 2017. To complement this initial set, we expand our dataset with an additional subset of 603 seized businesses, termed 'list B.' This information comes from the archives of the Italian Agency for the Administration and Destination of Seized and Confiscated Assets (whose Italian acronym is ANBSC). The main task of the ANBSC, a governmental authority, is the administration of all assets, including companies, seized from organized crime groups.

The dataset of seized and confiscated firms built so far, though useful in its own right, may not be considered exhaustive, since it reasonably includes only but a tiny fraction of the entire population of OC-linked firms. As reported in a study by the European Commission (Hulme et al., 2019), currently, only 1.1% of criminal profits from EU fraud are subject to confiscation across the European Union. Hence, there is a compelling rationale for seeking out other infiltrated firms through alternative data sources.

Thus we further extend our training sample with a third set of 1,575 firms, denoted as 'list C.' We extract this information from a commercial database based on press news, wherein we identify all companies with stakeholders or administrators implicated in OC-related legal proceedings in the period 2007 to 2017.

[5] According to Italian civil law, firms are allowed to consolidate certain variables in their financial statements if they satisfy at least two out of the following three conditions: 1) total assets are below 4.4 million euros; 2) revenues are below 8.8 million euros; 3) number of employees is less than 50. Companies must meet these criteria either in the first year of activity or in two consecutive years.

The majority of the infiltrated firms in our list were sourced from data compiled by the UIF, Italy's Financial Intelligence Unit. Established in 2007 within the Bank of Italy, the UIF is an independent authority tasked with preventing and combating money laundering and terrorist financing. In fulfilling its responsibilities, the UIF gathers financial data and information primarily through suspicious transaction reports (STRs) submitted by financial intermediaries, professionals, and other relevant entities. The information contained in STRs offers a valuable and extensive insight into transactions potentially associated with criminal activities. In 2023, the UIF received more than 150,000 STRs, which provide detailed descriptions of transactions considered suspicious by the reporting entities. The STRs undergo rigorous cross-referencing with extensive judicial and investigative records related to organized crime. A comprehensive search for individuals reported in STRs is conducted using databases maintained by the National Anti-Mafia Directorate (*Direzione Nazionale Antimafia e Antiterrorismo*, DNA). This meticulous process results in the creation of a confidential list of individuals who are under investigation for mafia-related crimes (or are reported in judicial documents connected to mafia-related crimes). Additionally, we include individuals for whom UIF has received information requests in connection with OC investigations, both from Italian investigative and judicial authorities and foreign Financial Intelligence Units. Consequently, we compile a fourth list of 27,029 companies (referred to as 'list D') with stakeholders or administrators linked to organized crime-related investigations. The accessibility of this unique and highly confidential roster substantially enhances the value of this study.

Integrating these various datasets allows us to leverage an extensive and robust selection of firms for our supervised learning approach, enabling a more nuanced understanding of the traits distinguishing infiltrated firms from their non-infiltrated counterparts. The combined count of infiltrated firms is 28,570, which does not precisely match the sum of the numbers in the four lists, due to some overlap (Table 1).

The way the years of infiltration are determined relies on the specific data source deployed. For firms in list A and B we do not know when they were infiltrated but we do know when they were seized, thus ceasing being infiltrated. In order to prevent results from being influenced by seizure-related operations or leaks or other news on the ongoing investigations, we only use data for all years up to the second before the seizure. For the years following the seizure, the data are excluded from the analysis since the firms are managed by judicial administrators, thus they cannot be considered as regular legal firms.

Table 1. Number of OC-linked firms used to train the model

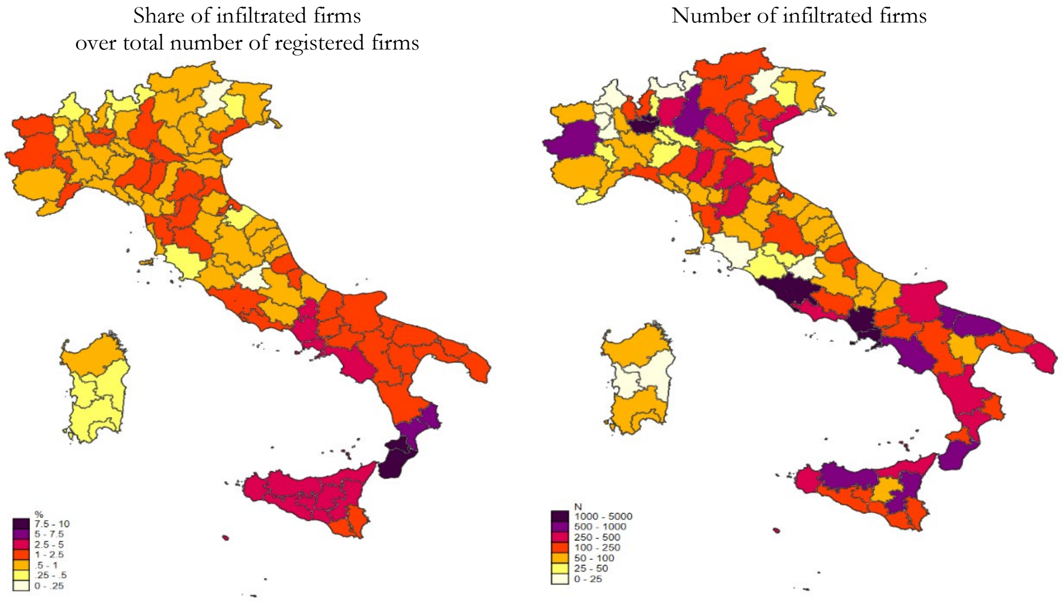
List A: businesses seized or confiscated by the judiciary as a result of the main anti-mafia investigations in the decade 2007 to 2017	229
List B: seized businesses drawn by the archives of ANBSC	603
List C: companies having stakeholders or administrators involved in OC-related legal proceedings derived from a commercial database	1,534
List D: companies having stakeholders or administrators involved, directly or indirectly, in OC-related investigations identified by the UIF based mainly on DNA data	27,029
Total number of unique infiltrated firms employed for the analysis	**28,570**

For lists C and D, we assume that the starting year of infiltration is the year where colluded stakeholders or administrators join the firms and we discard all data from previous years. Since we do not have a final year of infiltration for these cases, we use all data from the starting year of infiltration up to the last year of analysis, which is 2021.

Figure 1 depicts the geographical and sectoral distribution of infiltrated firms. Southern provinces display a higher share of infiltrated firms (over total registered firms at province-level), especially in Calabria (see the left panel of Figure 1), the region where 'Ndrangheta originated. Among Italian OC groups, 'Ndrangheta shows a greater inclination to investing in the legal economy (De Simoni, 2022) and this seems to be reflected in our sample. Looking at absolute numbers (see the right panel of Figure 1) shows that Milan, alongside Naples and Rome, is the city that hosts the largest number of infiltrated firms. Brescia, Turin and the region of Emilia-Romagna, also located in the North of Italy, show a significant number of firms connected to OC groups as well.

By examining the sectoral breakdown (Figure 2), it becomes apparent that OC groups tend to allocate their investments to sectors characterized by minimal skill prerequisites, strong reliance on the public sector, and a generally low emphasis on research and development (R&D). Indeed, the relative share of infiltrated firms in construction, water and waste, and entertainment is higher than the relative share of non-infiltrated firms in those sectors. Transportation and storage, a sector that is generally considered ancillary to many OC activities, shows also a relatively high concentration of infiltrated firms.

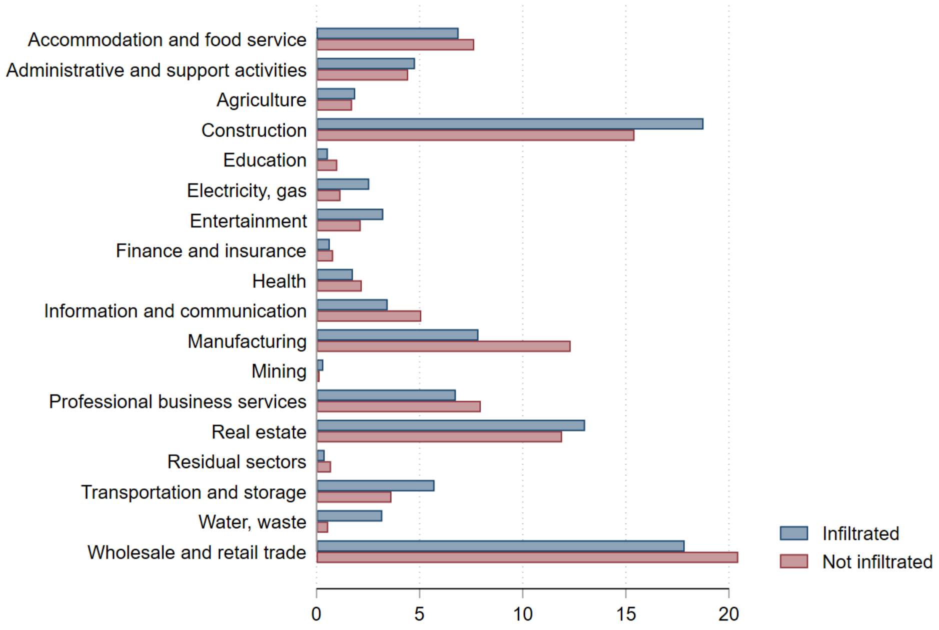
Figure 1. Geographical distribution of the sample of infiltrated firms[1]



(1) For the sake of simplicity, throughout the entire paper the terms OC-linked and infiltrated are used interchangeably, in order to identify the wider set of OC-linked firms.

Note: If a firm changes province of location over the years, we use the most recent available information. Some provinces (Barletta Andria Trani, Fermo, Monza Brianza and Sud Sardegna) are not covered in the National Official Business Register, as they have been recently formed. Firms from those provinces are enrolled in the Register of nearby provinces (Bari, Ascoli Piceno, Milano and Cagliari, respectively). For this reason, on the map, we arbitrarily allocate the same value to the missing provinces as that assigned to the province whose Register the firms are enrolled in (e.g. Monza Brianza has the same value of Milano).

Figure 2 Sectoral distribution of the sample of infiltrated firms[1]



(1) For the sake of simplicity, throughout the entire paper the terms OC-linked and infiltrated are used interchangeably, in order to identify the wider set of OC-linked firms.

Note: If a firm changes its economic sector over the years, we use the most recent available information.

4. The data

We built a unique dataset by merging different types of firm-level data spanning over 12 years, from 2010 to 2021. Four different types of data, each coming from distinct databases, are employed:

1. The firm-level financial statements come from the Cerved database, provided by the National Official Business Register, covering all limited liabilities, joint stock companies, and other companies legally mandated to deposit financial statements.
2. Firms' bank liabilities come from the Bank of Italy's Central Credit Register, where a customer is reported if outstanding debt to the intermediary is equal to €30,000 or more.
3. Employment data come from a database provided by the National Institute of Social Security (INPS).
4. Information pertaining to owners, boards of directors, and other firms' characteristics, such as location, sector of activity, and date of establishment, is sourced from the Chambers of Commerce.

A crucial feature of our research is the appropriate selection of the variables used to train the algorithm for detecting infiltrated businesses. Drawing from the most relevant papers in this field, we select a list of 32 variables and indicators that thoroughly characterize a firm's financial profile (see Table A1 in the Appendix). In particular, we focus our attention on financial variables and indicators that capture eight distinct features of a firm:

1. Five indicators measuring a firms' size along as many dimensions, such as total assets, revenues, equity, short-term liabilities and fixed assets;
2. Equity and liquidity, which are gauged based on seven indicators;
3. Indebtedness, measured by four distinct indicators that combine financial budget variables with firms' bank liabilities from the Bank of Italy's Central Credit Register;
4. Profitability: we use five different indicators extracted from the Cerved database;
5. Five indicators related to investment and cost structure;
6. Three budget indicators (cost of labour, revenues, added value) computed per labour unit;

7. Other elements of the financial statements: indicators such as inventory, accrued income, and liabilities;
8. Three opacity indicators at the firm-year level, computed at UIF and measuring opacity with respect to three different perspectives: ownership structure, administrators and other contributing factors (see Appendix B for definition and methodological details).

The legal form, the primary economic sector of activity, identified by the 2-digit NACE code, and the firm province location (110 provinces distributed in 20 regions in 2021) complete the set of features available for each firm.

Table A2 in the Appendix displays the main descriptive statistics of the variable used in our analysis by infiltration status. The dataset has been previously subjected to a very basic cleansing treatment in order to spot and resolve potential data inconsistencies; all monetary variables have been adjusted to 2021 constant prices.

Overall, the descriptive statistics confirm several key findings in the literature. Infiltrated firms tend to be larger, along all dimensions, including in terms of assets and revenues. However, higher revenues are not matched by higher profitability, as infiltrated firms generally perform worse according to profitability indicators. Another interesting finding concerns employment indicators: the cost of labour per employee is higher for firms linked to OC compared to the general economy. This descriptive evidence corroborates the results by De Simoni (2022). Regarding the lower level of profitability, it is noteworthy that firms linked to OC have a higher level of intermediate inputs and net purchases over revenues. Finally, infiltrated firms, on average, exhibit greater opacity than non-infiltrated firms.

Table A3 briefly shows the structure of the panel dataset by year.

5. Classification methodology

While classifying lawful and infiltrated firms might seem like a typical classification task, several important issues need consideration.

The initial concern revolves around the labelling process. Even though we select infiltrated firms by using the most current and trustworthy information, there is a possibility that some firms labelled as non-infiltrated might have connections to criminal organizations. This is a common challenge encountered in other research studies (Ravenda et al., 2015), and it does not have a straightforward solution. However, while in principle this could introduce a bias into our classifier, the large number of supposedly lawful firms makes this bias arguably negligible.

Another notable challenge in our classification task relates to the substantial geographic imbalance in the distribution of infiltrated firms, with a notable concentration in the Southern regions of Italy. While incorporating province dummy variables as predictors in our model indeed enhances its overall accuracy, owing to the strong link between geographic factors and infiltration, it also poses the risk of creating a model that is overly specialized for the Southern regions. This specialization might result in reduced effectiveness when assessing firms in the Northern regions, characterized by lower infiltration rates. In these regions, the model may become less sensitive to signs of infiltration, which could manifest differently. Nonetheless, accurately identifying OC-linked firms in regions historically less affected by mafia infiltration remains crucial. These businesses tend to receive less scrutiny, and detecting them could reveal previously undisclosed channels of OC infiltration.

To tackle this challenge, we apply a dual approach. We first develop a general model that in addition to the financial variables and indicators, includes the per-capita value added at the provincial level sourced from ISTAT. Importantly, we deliberately exclude additional geographical variables such as province or region dummies. The inclusion of per-capita value added, even though correlated with a firm's location, captures time-varying macroeconomic information that is likely associated with OC infiltration, as supported by previous research (Bernardo et al., 2021; Pinotti, 2015; Mocetti and Rizzica, 2021).

Then, in addition to the general model, we explore a more simplified model that entirely omits firm location and provincial macroeconomic data from consideration. This approach - i.e., the comparison of the results of the general vs. the restricted (simplified) model) - allows us to assess the impact of geographical factors on our classification task while providing insights into the model's performance without such inputs.

Another significant challenge arises from the substantial class imbalance in our dataset, with infiltrated firms constituting a very small fraction of the total firms. Specifically, the annual balance sheet records of firms labelled as infiltrated account only for 1.4% of the 11,426,981 total occurrences in our dataset. To address this imbalance, we employ a stratified under-sampling strategy with proportional allocation. Within this approach, non-infiltrated firms are stratified based on region and sector of activity. From each stratum, random samples are extracted in proportion to firms' distribution within the total population. This process is executed to ensure that, ultimately, the proportion of records for infiltrated firms equals approximately 50% of the sampled records used in our analysis.

This method serves to maintain the representation of the minority class by randomly removing instances from the majority class (non-infiltrated), thus achieving a balanced dataset. The advantage of this approach lies in its simplicity and its desirable additional capacity to alleviate the computational burden by downsizing the overall dataset. Moreover, under-sampling may help prevent the overfitting of the model to the majority class, promoting its suitability to address unseen data (generalizability). In this sense, the under-sampling process acts as a form of regularization, aiding in managing the model complexity and helping in the prevention of overfitting.

However, under-sampling has also its drawbacks. The observations from the majority class being dropped may contain valuable patterns contributing to defining the boundaries between the classes more efficiently: hence important information risks being lost. As a result, the classifier may be influenced by the specifics of the resulting sample. Additionally, if the minority class instances are not adequately representative of their class, thus causing the classifier to be somewhat biased in itself, under-sampling can exacerbate this bias, leading to poor generalizability.

We implement a range of strategies in order to address these particular concerns. One is deploying repeated sampling in order to enhance the robustness of the model's training process. Additionally, we made a deliberate decision to prioritize the overall generalizability of the model rather than solely focusing on achieving maximal performance levels.

Once the data engineering phase has been completed we applied the standard process for the development of a machine learning model, including the following steps:

a) Data Splitting
b) Feature Engineering
c) Model Selection
d) Model Calibration & Training
e) Model Evaluation

a) Data splitting

As an initial step, we divide the entire dataset into two sets with an 80/20 ratio for training and testing, respectively. The test set is essential for providing an unbiased evaluation of the model's performance on unseen data, which is crucial for assessing how the model might perform in real-world scenarios. Additionally, because the selection and fine-tuning of the model require a distinct dataset that is separate from the one used for model training, we create a validation set by extracting 20% of the total sample from the training dataset. This additional separation is vital for the effective validation and calibration of the machine learning model.

The partitioning of the dataset into training, validation, and test sets is carried out using the strategy of stratified sampling, where the stratum is the infiltration status. In a context of high imbalance between the classes, this approach ensures that the minority class is never under-represented. Furthermore, as we have economic and financial data observed over multiple years for each company, the stratified sampling is combined with cluster sampling. This means that all observations for the same company (cluster) are selected for each set, instead of individual observations. This ensures that each company is exclusively included in either the training or test set, preventing overlaps and, consequently, cross-contamination.[6]

After partitioning the dataset into training, validation, and test sets, the next step involves employing an under-sampling strategy to balance the two classes. As explained in the previous section, the stratified sampling procedure is applied, where strata are combinations of a company's region and sector of activity. In order to ensure that the model's performance is not solely attributed to the particular sample, we conduct the process five times in order to assess the consistency and reliability of the obtained results.

The entire splitting and under-sampling strategy is depicted in Figure 3, while the cardinality of the final datasets is showed in Table 2.

[6] See Zavitsanos et al. (2022) and Kapoor S. and Narayanan A. (2023).

Figure 3. Schema of the dataset splitting and under-sampling process

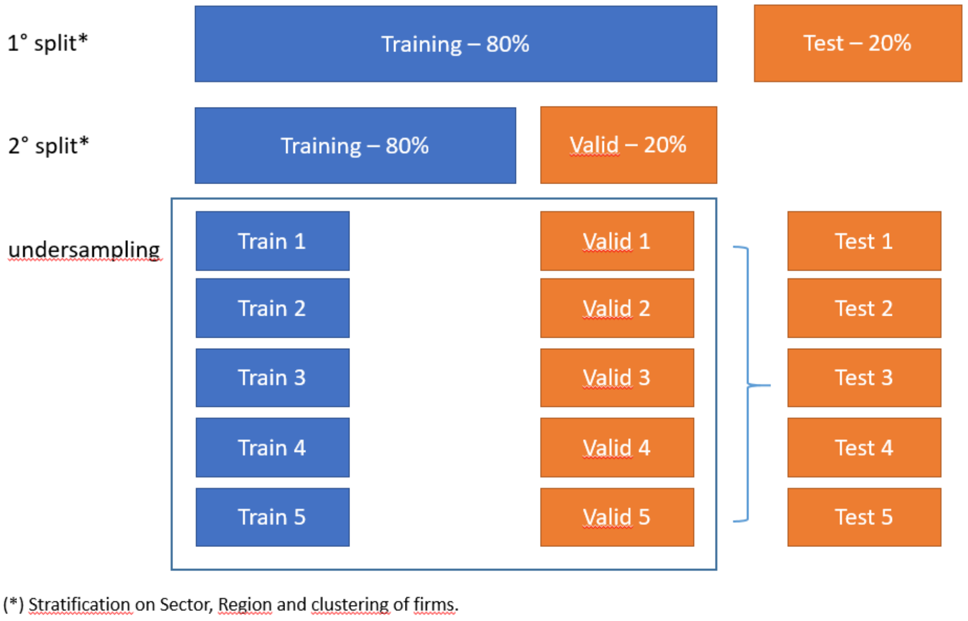


Table 2. Number of observations of train and test sets

Set	Training			Test		
	Total size	Non infiltrated	Infiltrated[1]	Total size	Non infiltrated	Infiltrated[1]
1	263,993	134,668	129,325	65,736	33,570	32,166
2	263,462	134,137	129,325	65,576	33,410	32,166
3	262,845	133,520	129,325	65,837	33,671	32,166
4	264,290	134,965	129,325	65,613	33,447	32,166
5	263,199	133,874	129,325	65,985	33,819	32,166

(1) For the sake of simplicity, throughout the entire paper the terms OC-linked and infiltrated are used interchangeably, in order to identify the wider set of OC-linked firms.

b) Feature engineering (Variable selection and transformations)

In the feature engineering step, we opt to include all available variables and indicators in the algorithm, except for the province dummies. We apply one-hot encoding to the categorical variable for the economic sector.[7] Additionally, we intentionally choose not to impute any missing data, which only affects the opacity indicators. This decision is made because XGBoost, the selected algorithm, is capable of effectively handling missing values.[8] Furthermore, missing data can be a distinctive characteristic of companies intentionally concealing information

[7] One-hot encoding is a method used in machine learning to transform a categorical variable in a set of *n* dummy variables (i.e. variables that can assume only ones or zeros values), where *n* is the number of distinct values of the original variable.

[8] "XGBoost supports missing values by default. In tree algorithms, branch directions for missing values are learned during training. Note that the gblinear booster treats missing values as zeros." - Frequently Asked Questions — xgboost 1.7.6 documentation – last accessed 4/8/23.

We also opt not to normalise numeric variables, as XGBoost and decision trees, in general, are capable of accommodating both binary and continuous features without being influenced by their scale, thus maintaining the integrity of their performance (Hastie et al., 2009). Unlike other machine learning algorithms, such as k-nearest neighbours or gradient descent-based methods, decision trees do not rely on distance-based calculations, making them inherently insensitive to the scale of numeric features. Since decision trees partition the feature space based on thresholds, the ordering and magnitude of the features do not affect their performance significantly.

c) Model selection

This phase aims first to select an appropriate machine learning algorithm based on the problem's unique attributes and requirements and, subsequently, to search for the optimal configuration that maximizes the model's performance.

As for the first objective, after an extensive search and experimentation with several algorithms, we select XGBoost, a highly efficient and scalable implementation of the boosting algorithm, granting a performance comparable to that of other state-of-the-art machine learning algorithms in most cases. XGBoost efficiently addresses the computational complexity and overfitting challenges often encountered by traditional boosting algorithms.

As for the second objective, ideally our classification end goal is to help the investigators by identifying as many infiltrated firms as possible. In this case, an important evaluation measure is the recall rate (also known as sensitivity or true positive rate). The recall rate measures the proportion of actual positives (infiltrated firms) that the model correctly identifies. A high recall rate indicates that the model can correctly identify most of the infiltrated firms. This may be crucial in financial intelligence or investigative scenarios as it allows the analyst (or the decision maker) to have a comprehensive view of the infiltrated firms in the given context under analysis and investigation, or in a given area or sector, etc., minimizing the chances of missing potentially infiltrated firms.

Conversely, our approach displays a higher degree of tolerance towards false positives, as they might signify a potential absence of information within the original dataset. This approach is based on the awareness that our sample cannot comprehensively encompass all conceivable illicit entities.

Another crucial metric, to evaluate the performance of the model, is precision, defined as the fraction of true positives over all the instances predicted as positive by the model. This is a key feature when the investigator (or analyst) has many subjects at his/her attention to be investigated or analysed, and the model can help to prioritize targets, i.e. allocate scarce resources to the riskiest targets.

Finally, we also computed specificity, defined as the ratio between true negatives and the total instances predicted positive. Appendix C provides an explanation of the performance metrics used, consistently with the common terminology applied.

d) Model calibration & training

To find the optimal combination of hyperparameters[9] that maximize the recall we use a randomized grids search with 50 random combinations of the parameters (Table 3).

Table 3. Combination of parameters used for search

Parameter	Values
Learning rate (Step size shrinkage)	[0.01, 0.05, 0.1, 0.2, 0.5]
N_estimators (Max num. of boosting trees)	[100, 500, 1000]
Max_depth (Maximum depth of a tree)	[3, 5, 10, 15]

We repeat the search for each training set and we compare their performances on the five validation sets. Table 4 shows the model's performance in terms of recall rate on the five validation sets. The recall scores exhibit a mean of approximately 73% for each validation set, with a narrow standard deviation of around 1%, thus indicating a consistent and stable performance when the parameters are altered.

Table 4. Variability of recall rate on validation sets

Validation set	Optimal parameter set	Max	Mean	St. dev.
valid1	[0.05, 1000, 15]	0.747	0.727	0.014
valid2	**[0.05, 500, 15]**	**0.755**	**0.735**	**0.014**
valid3	[0.05, 1000, 15]	0.754	0.734	0.014
valid4	[0.05, 1000, 15]	0.755	0.734	0.016
valid5	[0.05, 1000, 15]	0.753	0.732	0.015

Note: recall rate is computed at a cut-off point of 0.5.

In order to be conservative and prioritize the overall generalizability of the model rather than solely focusing on achieving maximal performance levels, we do not select the model that has the maximum recall, but pick the "second best" model, corresponding to

[9] In machine learning, a hyperparameter is a parameter whose value is used to control the learning process.

the 2nd validation set, with the following settings: learning rate set at 0.05, 500 decision trees (n_estimators) and a maximum depth of 15 levels. Other parameters are left to default values. The trained model shows a recall rate of 75.5% and a precision rate of 75.0% on the second under-sampled validation set (Table 5).

Table 5. Evaluation of model's performance on validation sets

Validation set	Recall	Precision	Specificity
valid1	0.747	0.745	0.752
valid2	**0.755**	**0.738**	**0.741**
valid3	0.754	0.744	0.746
valid4	0.755	0.732	0.732
valid5	0.753	0.742	0.746

Note: recall, precision and specificity are computed at a cut-off point of 0.5.

e) Model evaluation

This phase aims to evaluate the trained model on the test sets to obtain unbiased performance metrics, analyse the results and assess the model's effectiveness. Our achieved performance closely aligns with that obtained on the validation set, reinforcing the confidence that the models are not overly tailored to the training data, thereby mitigating concerns of overfitting. Moreover, all the metrics considered show quite stable values across the sets, suggesting that the capacity of the model of making correct predictions, both positive and negative, is robust (Table 6).

Table 6. Evaluation of model's performance on test sets

Test set	Recall	Precision	Specificity
test1	0.752	0.737	0.743
test2	**0.756**	**0.738**	**0.742**
test3	0.755	0.740	0.747
test4	0.756	0.737	0.740
test5	0.755	0.738	0.746

Note: recall, precision and specificity are computed at a cut-off point of 0.5.

Our findings align with analogous research in the existing body of literature that leverages annual financial statements. In particular, Ravenda et al. (2015) reported a sensitivity of about 76% by using a logistic model approach.[10]

To appreciate the informative gain guaranteed by our model, it must be taken into account that we are trying to intercept a condition that rarely occurs within the entire

[10] It needs stressing that in Ravenda et al. (2015) the sample of infiltrated firms is only 852 units and model performance is evaluated on the same set used to estimate the model.

population of firms and which is therefore difficult to detect, thus showing a remarkable enhancement compared to predictions solely driven by the toss of a coin[11] (Table 7).

Table 7. Comparison with 'baseline' models

Model	Accuracy	Recall	Specificity
XGB	0.743	0.756	0.742
Stratified	0.501	0.490	0.511
Most frequent (0)	0.986	0.000	1.000
Uniform	0.500	0.499	0.500
Less Frequent (1)	0.014	1.000	0.000

We also investigated the performance of the model at various cut-off points, by comparing the different values of recall (sensitivity), precision and specificity.[12] At a cut-off of 0.5 we achieve a near-optimal balance between sensitivity and specificity. To achieve a sensitivity of 80%, it is necessary to reduce the cut-off threshold to about 0.43. However, this comes with a sensitivity loss of 5 percentage points, indicating that the model may miss more true positive cases, and a precision loss of more than 2 percentage points, meaning there may be more false positive cases. Conversely, if we increase the cut-off to 0.6, the sensitivity drops to 68.3%, meaning the model is less effective at identifying true positive cases. However, the specificity increases to 79.8% and precision gains almost 3 percentage points. Variations in the cut-off values can be taken into account during the operational use of the model to broaden or limit the number of potential infiltration cases reported by the model, thus calibrating potential false positives and false negatives accordingly (Figure 4).

[11] The baseline models make predictions based only on target distribution, ignoring the input features. We use 'Dummy Classifiers' from Scikit Learn library. For further details, see https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html .

[12] Values are computed on validation set, as cut-off selection can be viewed as part of the overall calibration of the model parameters.

Figure 4. Performance metrics across cut-off points

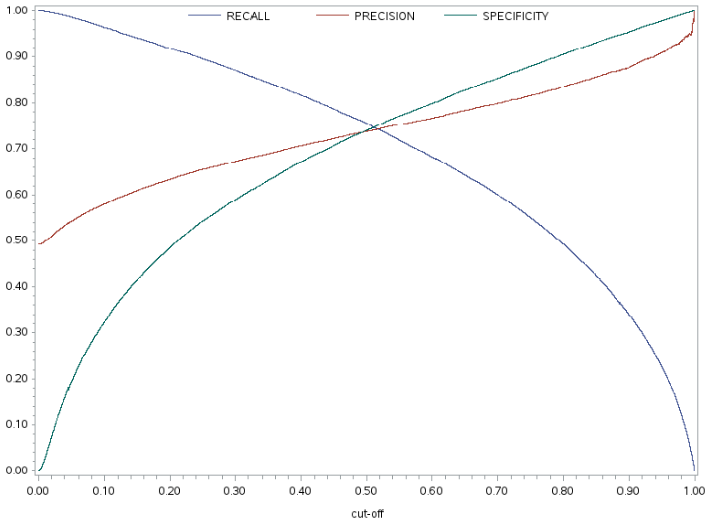


Table 8 highlights that when the cut-off point is pushed to extreme values, such as 0.95 or 0.99, the model effectively identifies businesses that are allegedly lawful in over 95% of instances. However, this strategic adjustment comes with a notable trade-off in the form of a drastic reduction in recall.

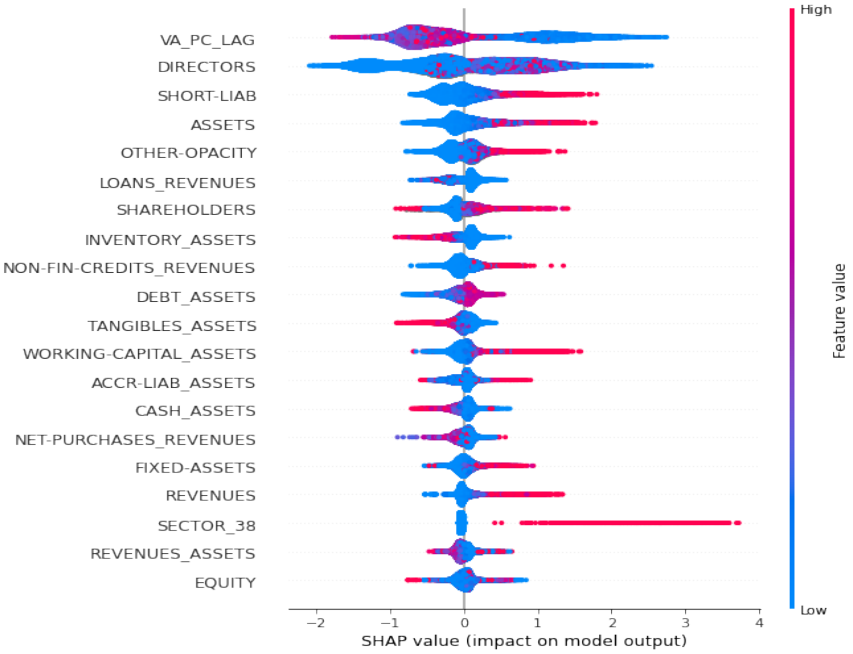
Table 8. Performance metrics with high cut-off points

Cut-off point	Recall	Specificity
0.80	0.493	0.905
0.85	0.426	0.931
0.90	0.339	0.954
0.95	0.225	0.978
0.99	0.071	0.996

To single out the primary factors influencing the model's output, we employ the SHAP (SHapley Additive exPlanations) framework. This methodology, grounded in game theory, offers a comprehensive approach for explaining the output of various machine learning models. (Lundberg and Lee, 2017). The model predictions are recalculated adding or changing a variable's value and see how this affects the performance. This approach allows us to discern the most and least impactful features on the model, as well as the extent of their influence, whether positive or negative.

Overall, the findings highlight the highest level of importance attributed to province-level per capita value added (*va_pc_lag*) and the opacity of directors (*directors*). As expected, the results reveal that in provinces with lower per capita value added, there is a higher probability of detecting infiltrated firms. Following in significance are a group of size-related variables: as a firm's size, measured by assets (*assets*), revenues (*revenues*), and short liabilities (*short_liab*), increases, so does the probability of predicting infiltration. The other two opacity indicators (*other-opacity and shareholders*) also exhibit a noticeable influence on the model's predictions. Debt-related variables (*loans_revenues* and *debt_assets*) show a moderate level of importance. Additionally, investment indicators, specifically tangible assets over total assets (*tangibles_assets*), and liquidity indicators, namely cash over total assets (*cash_assets*), have a negative impact on the likelihood of model prediction (Figure 5).

Figure 5. Summary plot of SHAP values - top 20 variables



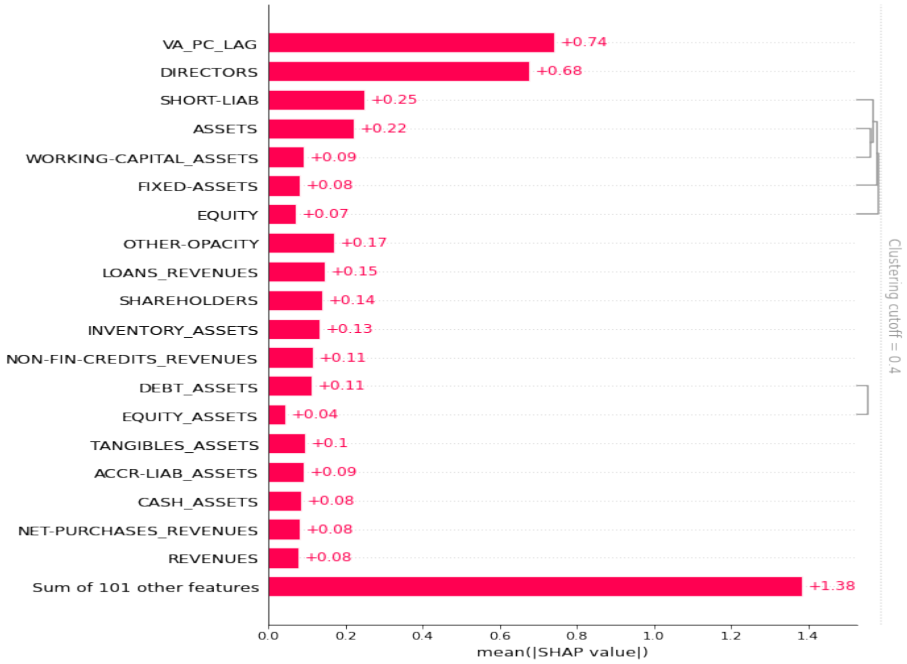
For the model training we made the deliberate choice to retain all the variables/indicators as identified and described in Section 4, even in instances where significant correlations exist among them. This approach was adopted due to the inherent resilience of decision trees to multi-collinearity. In our specific case, retaining all variables resulted in a marginal improvement in predictive performance. However, it is noteworthy that the presence of highly interrelated variables has an impact on the calculation of variable importance.[13] This arises from the fact that highly correlated variables can be deployed interchangeably during the node splitting process.[14]

[13] Importance provides a score that indicates how often each feature was used in the construction of the boosted decision trees within the model.

[14] For example, an article found at https://vishesh-gupta.medium.com/correlation-in-xgboost-8afa649bd066 empirically found that, starting from a model that have a variable var1 with a certain importance i1 and adding a perfectly correlated new variable var1_new, the new model has same performances, but different variables' importance, respectively i1' e i1_new , with i1' < i1 and sum(i1', i1_new) > i1.

To overcome this obstacle, we use a direct approach that measures feature redundancy through model loss comparisons directly during SHAP computation.[15] Typically, this results in much more accurate measures of feature redundancy than using an external unsupervised method like correlation. Using this approach and recalculating the SHAP values as in Figure 6, we confirm that per capita province-level value added and opacity of directors are the most influential factors for the model, followed by the dimensional variables group plus working capital over assets, whose contribution must be considered overall, in light of the close interrelationship between them.

Figure 6. Bar plot of SHAP values - top 20 variables with clustering



6. Robustness analysis

Since we are dealing with a potential bias due to the substantial geographic imbalance in the distribution of infiltrated firms in favour to Southern regions of Italy, it is important to perform a region-wise evaluation of the model by assessing the model's performance separately by geographic area. For this purpose, we consider the conventional division of the Italian territory into its four distinctive areas: North-West, North-East, Centre, South and Islands. Thus, we conduct a separate evaluation of recall, precision and

[15] We use shap.utils.hclust method, that builds a hierarchical clustering of the feature by training XGBoost models to predict the outcome for each pair of input features. See https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/bar.html#Using-feature-clustering for further details.

specificity for each of the distinct geographical areas, resulting in the metrics showed in Table 9.

The model's efficacy diminishes in regions characterized by a scarcity of positive instances available for training, namely in the North and Centre of Italy. These findings signal the need that another model omitting any geographical covariate be developed, so as to assess to what extent performance declines and whether categorizing firms exclusively based on the financial variables and indicators is feasible. By eliminating the per-capita province-level value added and subsequently retraining the model under the same set of parameters, the ensuing analysis reveals a foreseeable reduction in model performance.

Table 9. Test set performance by geographical breakdown

	With per-capita province-level value added			Without any geographical variable		
	Recall	**Precision**	**Specificity**	**Recall**	**Precision**	**Specificity**
North-West	0.650	0.692	0.805	0.737	0.645	0.727
North-East	0.624	0.644	0.830	0.713	0.593	0.759
Centre	0.686	0.730	0.782	0.742	0.686	0.707
South and Islands	0.858	0.775	0.582	0.711	0.795	0.692
Total	**0.756**	**0.738**	**0.742**	**0.723**	**0.712**	**0.719**

In response to a decline of three percentage points in recall, precision and specificity, the model's performance in the absence of the province variable demonstrates a greater uniformity across the four distinct geographical areas. This observation suggests a promising future research for advancing the model by exploring alternative factors, transformations, or combinations of variables that are not directly contingent on geographical location.

7. Risk score computation and applications for AML and prudential supervision

One of the most desirable outcomes of our model is its ability to compute a risk score, which can be interpreted as the probability of infiltration, for the whole population of Italian registered capital companies. Table 10 shows the frequency distribution of the estimated risk score for 931,163 firms based on the most recent available year, i.e. 2021: 78.4 per cent of firms are low risk, having a risk score of less than 0.5; the remaining 21.6 per cent of firms are labelled as risky businesses according to our model, even though, only 1.8 per cent of them are to be considered as very high-risk, having a risk score of more than 0.95.

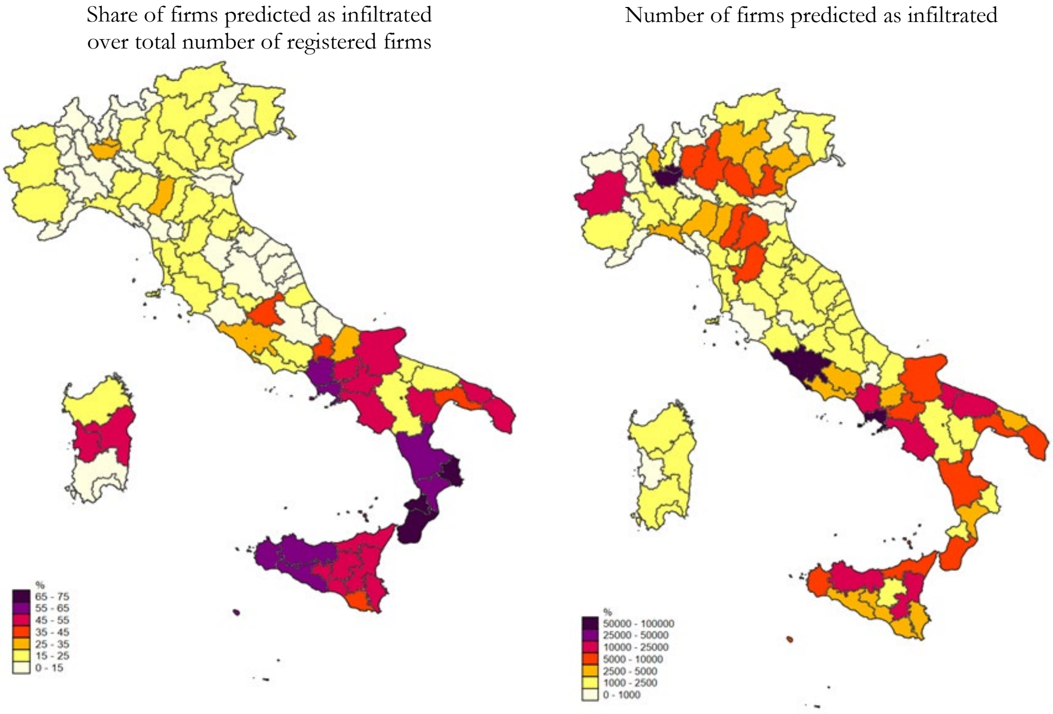
Table 10. Frequency distribution of estimated risk score - year 2021

Risk score	N	%
Up to 0.5	729,433	78.4
From 0.5 to 0.8	129,799	13.9
From 0.8 to 0.95	54,969	5.9
From 0.95 to 0.99	13,916	1.5
Over 0.99	3,046	0.3
Total	**931,163**	**100.0**

Note: Per-capita firm-level value added is included among the covariates of the model.

Figure 7 shows the provincial map of firms predicted as infiltrated by the model, both in terms of share over total number of registered firms in the province (left panel) and absolute number of firms (right panel). As expected, the southern provinces exhibit a relatively high proportion of firms predicted as infiltrated, reaching peaks in select provinces of Sicily, Calabria, and Campania. Nonetheless, the provinces of Naples, Rome, Milan, and Brescia stand out with the highest absolute count of firms projected to be associated with infiltration.

Figure 7. Geographical distribution of firms predicted as infiltrated (score higher than 0.50) Year 2021



Share of firms predicted as infiltrated over total number of registered firms

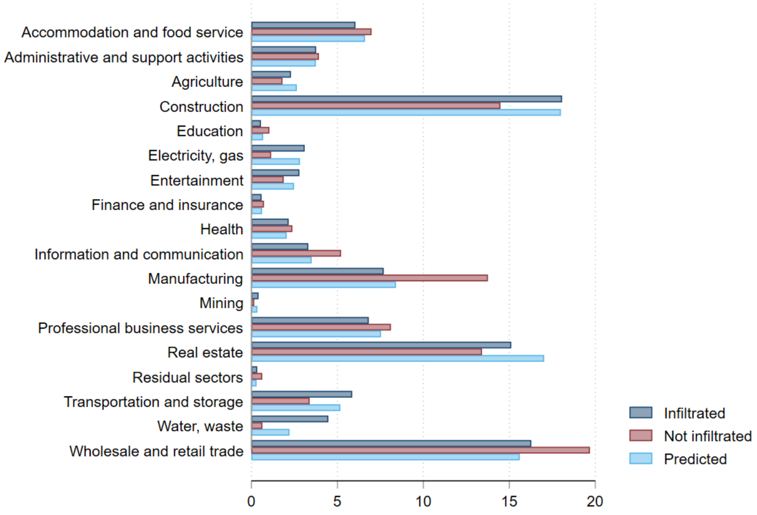
Number of firms predicted as infiltrated

Note: The most recently formed provinces of Barletta Andria Trani, Fermo, Monza Brianza and Sud Sardegna, are not covered in the National Official Business Register. Firms from those provinces are assigned to the Register of Bari, Ascoli Piceno, Milano and Cagliari, respectively. For this reason, on the map, we allocate the same value to the missing provinces as that assigned to the province where the firms are categorized in the Register. (e.g., Monza Brianza has the same value of Milano). Per-capita firm-level value added is included among the covariates of the model.

As for sectoral distribution, Figure 8 displays the proportion of firms predicted as infiltrated within each of the 21 NACE level 1 sectors (represented by light blue bars), revealing a broad conformity with the distribution of the infiltrated firms in the training sample (depicted by dark blue bars). Notably, the real estate sector diverges as the model predicts a comparatively greater percentage of infiltrated firms compared to the training sample. This highlights how there may be a need for further investigation in this sector in relation to its increasing role in illegal activities and money laundering processes, especially considering the significant flow of public funds disbursed under various stimulus public programs.

Contrary to expectations, the waste and water sector demonstrates a lower percentage of predicted infiltrated firms compared to the training sample, despite the fact that the corresponding dummy variable holds significant importance in the construction of the model.

Figure 8. Sectoral distribution of firms predicted as infiltrated (score higher than 0.50) Year 2021



Note: Per-capita firm-level value added is included among the covariates of the model.

The risk score associated to the Italian registered limited liability companies has several potential applications for AML purposes. It can serve as an additional red flag indicator for UIF institutional functions, when used in conjunction with additional information regarding potential OC connections. The score can be also computed as an aggregate risk indicator both at a geographical or sectoral level (as showed in Figures 7 and

8), which may provide interesting insights, particularly for UIF strategic analysis and within the National Money Laundering Risk Assessment.[16]

For prudential supervision purposes, a potential application of the indicator can be obtained by computing the financial exposure of each banking institution towards risky companies (i.e., the ones with a high risk score). Such an indicator may provide useful information for risk-based prudential oversight.

In a broader context, one could envisage leveraging the model's predictions by Financial Intelligence Units (FIUs) and Law Enforcement Agencies (LEAs) to prioritize and streamline investigative efforts, particularly in cases where a multitude of entities necessitate scrutiny during the planning phase of investigative activities, or during monitoring of public funds.

8. External validation

In order to assess the model's real-world performance, an external validation has been conducted with independent data on two significant lists of firms. The first list of businesses includes a selected sample of companies that are in the so-called whitelists of the Italian Prefectures. These whitelists are established at each Prefecture with the primary objective of enhancing the effectiveness of anti-mafia assessments for business activities deemed to be at a higher risk of mafia infiltration. Certain specific categories of businesses are obligated to be registered on the whitelist to engage in direct or indirect contracts with the public administration. The Prefecture is granted a 90-day period for approving registrations on the whitelist. In cases where there is no response within the timeframe, contracting authorities still have the authority to proceed with contract execution. It is worth noting that there are several potential gaps or shortcomings in these lists: for instance, companies can be automatically included if the Prefecture fails to respond within the specified timeframe. For these reasons, even though a significant majority of the businesses registered in the whitelists can be considered mafia-free, there might be still a percentage of firms that could be potentially linked to OC.

The second list is the list of firms with "interdittiva antimafia" (anti-mafia injunction), which are the businesses that are subject to restrictions or prohibitions imposed by authorities due to alleged criminal connections or mafia influences. Overall in

[16] E.g., see National money laundering and terrorist financing risk assessment report 2018 available at https://www.dt.mef.gov.it/en/pubblicazioni/analisi_nazionale_rischi_riciclaggio/.

Italy local Prefectures adopt nearly 700 antimafia injunctions on average every year, with a substantial increase, exceeding 1,000, in both 2020 and 2021, coinciding with the outbreak of the COVID-19 pandemic (Ministero dell'Interno, 2023).

While a company subject to an "interdittiva" can be assessed as (highly) likely to be infiltrated, the inclusion in such list by itself does not represent, of course, conclusive evidence of infiltration (a circumstance that can only be ascertained at investigative or judicial level), due for example to the possibility of injunctions being subject to review and revocation, which makes the list prone to errors.

While the whitelist and "interdittiva" represent distinct legal provisions, it can be asserted that the eligibility criteria for denying registration on the whitelist are equivalent to those for being subject to an "interdittiva", as both measures share the same rationale and purpose of safeguarding economic public order, promoting fair competition among businesses, and ensuring the proper functioning of public administration. The key difference between the two is that the former is triggered at the request of a firm engaging in a contract with the public administration, while the latter is triggered by authorities, typically in connection with some law enforcement action.

The list of firms subject to "interdittiva" comprises a sample of 1,667 capital companies drawn from the lists of all 103 Italian Prefectures. The numbers of the firms included in the whitelists are much bigger, in the order of hundreds of thousands for the whole Italy, and we have been provided with the data of four provinces (Reggio Calabria, Rome, Reggio Emilia and Turin), for a total of 23,248 capital companies. It is worth noting that these provinces are particularly significant due to their high incidence of mafia infiltration, either 'native', as in the case of Reggio Calabria, or 'imported', as in the other three provinces.

Table 11 presents the key findings regarding our model's performance on both lists of companies. Focusing on the firms located in the four selected provinces, a first noticeable result is that the mean score of firms subject to "interdittiva" is much higher (roughly, 70% higher), than the mean score of whitelist firms (0.594 vs. 0.346). The difference is even starker when the median is considered: that for the former sample is more than twice as much as the median of the latter sample (0.657 vs. 0.249). We also conducted statistical tests to assess whether the average scores of firms subject to "interdittiva" are significantly higher than those of whitelist firms, and the result shows a difference which is statistically significant at a 1 percent confidence level, thus reaffirming the model's capacity to discriminate between the two categories of firms. Encouragingly, the mean score for firms with "interdittiva" located in all Italian provinces is even higher than the corresponding figure for the four provinces (0.626).

Analogously, the percentage of firms predicted as infiltrated for the firms subject to "interdittiva" (61.2%, when focusing on the four provinces) is more than twice as much as the corresponding figure for whitelist firms (29.5%). Arguably, such figures appear roughly consistent with the *in-sample* results obtained in terms of recall and specificity.

Table 11. Main results of the external validation

	Firms with "interdittiva antimafia"		**Firms included in whitelists**
	All sample	**Provinces of Reggio Calabria, Rome, Reggio Emilia and Torino**	**Provinces of Reggio Calabria, Rome, Reggio Emilia and Torino**
	(A)	**(B)**	**(C)**
Number of firms	1,667	322	23,248
Mean of the score	0.626	0.594	0.346
Median of the score	0.710	0.657	0.249
Percentage of firms predicted as infiltrated (cut-off at 0.5)	64.6	61.2	29.5
		Test on the difference between means	
		B > C	
Statistic test t		12.10	
Significance		***	

Some interesting results on the model's ability to identify infiltrated firms across different areas of the country (see discussion above, in Section 7) emerge from the evidence disaggregated by province (Table 12). The model exhibits a lower rate of false negatives (i.e., companies with "interdittiva" that are predicted as non-infiltrated) in the provinces of Reggio Calabria and Rome, compared to the two northern Italian provinces. This is due to the model's superior recall performance in provinces where a larger share of positive cases is used to train the model. Conversely, for the same reason, the model records a lower rate of false positive (i.e., companies in the whitelist predicted as infiltrated) in the northern provinces of Reggio Emilia and Turin.

Table 12. Results of the external validation by province

	Reggio Calabria	Rome	Reggio Emilia	Torino	All four provinces
			Firms with "interdittiva"		
Number of firms	131	56	62	73	322
False negatives rate (%): firms predicted as non-infiltrated	28.2	23.2	61.3	50.7	38.8
			Firms in whitelists		
Number of firms	752	11,861	3,148	7,094	22,855
False positives rate (%): firms predicted as infiltrated	50.7	34.5	19.3	21.8	29,0

9. Conclusions

In this study we develop a Machine Learning algorithm in order to detect legally registered firms potentially connected to organized crime.

To this end, a sample of Italian capital firms highly likely to be linked to OC is compiled by resorting to four different lists of firms from various sources, including law enforcement seizures, the archives of the Italian Agency for the Administration and Destination of Seized and Confiscated Assets, and a commercial database. In addition to these, our primary data source is the UIF's own archives, allowing us to identify a list of firms with stakeholders and administrations reported in Suspicious Transaction Reports (STRs) and involved, directly or indirectly, in investigations related to OC. This led to the creation of a list comprising over 28 thousand companies connected to OC investigations with a high probability.

We employ a highly varied list of financial and budget indicators and variables, identified on the basis of the latest literature on criminal infiltration in real economy. The 32 variables are computed by using several different sources.

We leverage this comprehensive and innovative dataset to construct a machine learning classifier with the ability to detect legally established entities that could potentially be connected to OC. For this purpose, we employ an XGBoost (eXtreme Gradient Boosting), a state-of-the-art machine learning algorithm. The sample of firms labelled as OC-linked are compared with stratified random samples of alleged lawful firms in order to train and test the model. The main output of the algorithm is a risk score computed for the whole population of registered capital companies.

The ML algorithm successfully identifies approximately 76% of the OC-linked firms (recall) and 74% of the allegedly lawful firms (specificity).

The model's performance has been also assessed in a real-world scenario using independent data, which includes two lists of firms: the whitelists of the Italian Prefectures for enhancing anti-mafia assessments for businesses contracting with the public administration, and firms with "interdittiva antimafia" (anti-mafia injunction), subjected to restrictions due to alleged criminal or mafia ties. The results obtained by computing our algorithm on such firms are encouraging on the model's potential for practical applications.

The risk score associated to Italian registered limited liability companies has multiple possible uses for AML purposes, including functioning as an additional red flag indicator for UIF institutional activities. The score can also be calculated as an overall risk indicator at either a geographical or sectoral level. This could offer valuable insights, especially for strategic analysis by UIF and the National Money Laundering Risk Assessment. It might also contribute to identify intermediaries more exposed to risky companies, in risk-based prudential oversight. Looking at the bigger picture, one could consider using the model's predictions to help Financial Intelligence Units (FIUs) and Law Enforcement Agencies (LEAs) prioritize and streamline investigative efforts. This would be particularly beneficial, for example, in the monitoring of public funds.

There are appear to be several opportunities for future research. First of all, we could adopt a more extensive approach in the data preparation phase. For example, we could try to reduce the influence of extreme values applying trimming or winsorization tools, and testing the use of algorithms for identifying and removing outliers. Secondly, we intend to explore the use of alternative supervised machine learning algorithms, like Random Forests, CatBoost and Neural Networks. Finally, an alternative avenue to explore could involve adopting a dynamic approach that capitalizes on the evolving dynamics of financial variables within each firm. This approach would diverge from the conventional practice, which we adopt, of considering yearly occurrences in isolation. However, an approach of this type involves a significant reduction in the number of units for training, given that several companies do not have a complete historical series of financial statements for the period observed.

Appendix A

Table A1. List of financial variables/indicators

Dimension of analysis	Variable/indicator	Abbreviation	Source
Size	Assets	ASSETS	Central business registry
	Revenues	REVENUES	
	Equity	EQUITY	
	Short term liabilities	SHORT-LIAB	
	Fixed assets	FIXED-ASSETS	
Equity and liquidity	Cash over assets	CASH_ASSETS	Central business registry
	Equity over assets	EQUITY_ASSETS	
	Short-term assets over short-term liabilities	S-ASSETS_S-LIAB	
	Revenues over assets	REVENUES_ASSETS	
	Cash flow	CASHFLOW	
	Non-financial credits over revenues	NON-FIN-CREDITS_ REVENUES	
	Working capital over assets	WORKING-CAPITAL_ASSETS	
Indebtedness	Leverage (granted loans over equity)	LEVERAGE	Central business registry Central Credit Registry
	Granted loans over revenues	LOANS_REVENUES	
	Net debt (granted loans - cash) over EBITDA	NET_EBITDA	
	Total debt over assets	DEBT_ASSETS	
Profitability	EBITDA over revenues	EBITDA_REVENUES	Central business registry
	EBITDA over assets	EBITDA_ASSETS	
	ROI	ROI	
	ROE	ROE	
	ROA	ROA	
Investment (internal vs external resources) and cost structure	Tangibles over assets	TANGIBLES_ASSETS	Central business registry
	Cost of rents and leases over revenues	COST_REVENUES	
	Net purchases over revenues	NET-PURCHASES_ REVENUES	
	Intermediate inputs over revenues	INTERM_ REVENUES	
	Capital expenditure	CAP-EXP	
Employment	Cost of labour over number of employees	LABOUR_EMPL	Central business registry National Institute for Social Security database
	Revenues over number of employees	REVENUES_EMPL	
	Added value over number of employees	VALUE_EMPL	
Other elements	Accrued liabilities over assets	ACCR-LIAB_ASSETS	Central business registry
	Accrued incomes over assets	ACCR-INCOMES_ ASSET	
	Inventory over assets	INVENTORY_ ASSETS	
Opacity	Opacity of shareholders	SHAREHOLDERS	Chambers of commerce registry
	Opacity of directors	DIRECTORS	
	Other elements of opacity	OTHER-OPACITY	

Table A2. Descriptive statistics of the dataset

		Infiltrated firms[1]			Non-infiltrated firms		
Dimension of analysis	Variable/indicator	Mean	St. Dev.	Missing items (%)	Mean	St. Dev.	Missing items (%)
Size	Assets	4,906	9,137	0.00	1,793	4,735	0.00
	Revenues	2,921	6,775	0.00	1,277	3,812	0.00
	Equity	1,264	3,161	0.00	531	1,795	0.00
	Short term liabilities	2,244	4,186	0.00	792	2,145	0.00
	Fixed assets	1,894	4,137	0.00	685	2,165	0.00
Equity and liquidity	Cash over assets	0.113	0.197	0.00	0.159	0.2203	0.00
	Equity over assets	0.076	1.004	0.00	0.141	0.9153	0.00
	Short-term assets over short-term liabilities	4.493	15.363	0.00	4.477	14.4719	0.00
	Revenues over assets	0.793	1.175	0.00	0.977	1.1214	0.00
	Cash flow	49.386	703.267	0.00	20.936	385.840	0.00
	Non-financial credits over revenues	25.212	76.021	0.00	9.567	43.077	0.00
	Working capital over assets	2,590	4,988	0.00	1,000	2,673	0.00
Indebtedness	Leverage (granted loans over equity)	2.690	11.456	0.00	2.698	10.286	0.00
	Granted loans over revenues	17.074	101.176	0.00	10.523	73.882	0.00
	Net debt (granted loans - cash) over EBITDA	0.859	23.328	0.00	1.128	20.515	0.00
	Total debt over assets	0.862	0.937	0.00	0.791	0.869	0.00
Profitability	EBITDA over revenues	-2.188	8.297	0.00	-1.003	5.427	0.00
	EBITDA over assets	0.020	0.231	0.00	0.044	0.250	0.00
	ROI	0.121	0.927	0.00	0.169	0.870	0.00
	ROE	0.038	1.117	0.00	0.072	1.054	0.00
	ROA	-0.043	0.286	0.00	-0.033	0.287	0.00
Investment (internal vs external resources) and cost structure	Tangibles over assets	0.269	0.397	0.00	0.320	0.423	0.00
	Cost of rents and leases over revenues	3.183	10.048	0.00	1.701	6.461	0.00
	Net purchases over revenues	0.432	1.582	0.00	0.404	1.302	0.00
	Intermediate inputs over revenues	2.570	7.325	0.00	1.591	4.842	0.00
	Capital expenditure	-100.588	324.551	0.00	-40.795	184.309	0.00
Employment	Cost of labour over number of employees	19.736	24.268	0.00	17.912	20.657	0.00
	Revenues over number of employees	357.243	700.113	0.00	215.453	433.541	0.00
	Added value over number of employees	54.503	124.085	0.00	41.199	82.276	0.00
Other elements	Accrued liabilities over assets	0.012	0.040	0.00	0.012	0.039	0.00
	Accrued incomes over assets	0.011	0.039	0.00	0.012	0.038	0.00
	Inventory over assets	0.147	0.273	0.00	0.154	0.262	0.00
Opacity	Opacity of shareholders	23.488	30.069	12.33	13.582	26.084	14.30
	Opacity of directors	33.293	41.330	18.62	19.546	37.304	22.57
	Others elements of opacity	25.536	47.259	7.02	15.625	35.887	7.73

(1) For the sake of simplicity, throughout the entire paper the terms OC-linked and infiltrated are used interchangeably, in order to identify the wider set of OC-linked firms.

Note: mean and standard deviation are computed on winsorized data at the 1st and 99th percentiles.

Table A3. Structure of the panel dataset

Year	Infiltrated firms[1]	Non-infiltrated firms	% of infiltrated[1]
		Number of records	
2010	13,231	894,738	1.48
2011	13,661	906,929	1.51
2012	13,668	902,023	1.52
2013	13,581	904,407	1.50
2014	13,690	914,611	1.50
2015	13,887	932,143	1.49
2016	13,917	942,194	1.48
2017	13,956	958,721	1.46
2018	14,073	985,884	1.43
2019	13,918	1,006,100	1.38
2020	13,001	997,485	1.30
2021	10,908	920,255	1.19
Total	**161,491**	**11,265,490**	1.43
		Number of firms	
Total	**28,570**	**1,804,278**	**1.58**

(1) For the sake of simplicity, throughout the entire paper the terms OC-linked and infiltrated are used interchangeably, in order to identify the wider set of OC-linked firms.

Appendix B
The opacity indicators

We use the term "opacity" to encompass various methods used to conceal the true owner of a firm, such as employing trusts, using different jurisdictions or intermediaries. Existing extensive research has linked a lack of transparency to a higher likelihood of engaging in fraudulent activities (FATF, 2018).

Originating from an internal parallel project within the UIF, we compute three comprehensive opacity indicators on a firm-year basis: one appraises owner opacity, another assesses director opacity, and the third examines supplementary contributing factors. Using data sourced from the Italian Chamber of Commerce database (Infocamere), we gather information encompassing firms' characteristics (such as location, address, legal structure) as well as details pertaining to their proprietors and directors.

Owner opacity is constructed from 18 fundamental indicators related to owner characteristics. These indicators include metrics such as unusual distribution of firm shares among owners, consideration of foreign shareholders' risk based on their country of origin, categorization of owners as physical individuals or legal entities, presence of very young or elderly owners, identification of high-risk owner types (e.g., trusts or foundations), and the turnover rate of owners. We aggregate these fundamental indicators to establish the comprehensive indicator using a weighted average approach. Each fundamental indicator is assigned a weight based on the relative difficulty of obtaining the underlying data.

The same rationale is applied to build the opacity indicator for directors, albeit employing half the number of elementary indicators. This includes considerations of foreign and non-physical directors, as well as those who are very young or advanced in age. Factors such as turnover and distinct designations, like trusts and foundations, are also factored in. Additionally, we introduce a sub-indicator for identifying potential figureheads, quantified by the instances where a director holds the same position in multiple firms.

The third indicator captures miscellaneous attributes unrelated to shareholders and directors that could augment a firm's opacity. This encompasses scenarios like multiple firms sharing an identical address. Furthermore, we monitor the frequency of alterations in legal status, denomination, location, and sector of activity over a 5-year interval. Similar to the owners' indicator, we apply analogous principles for the weighting and aggregation of the directors' and other elements' opacity indicators.

Appendix C

Terminology and model performance metrics

True positive (TP): a model result that correctly indicates the presence of a condition (e.g. infiltration).

True negative (TN): a model result that correctly indicates the absence of a condition (e.g. non-infiltration).

False positive (FP), Type I error: a model result which wrongly indicates that a particular condition is present.

False negative (FN), Type II error: a model result which wrongly indicates that a particular condition is absent.

Metric	Description	Formula
Recall (Sensitivity or True positive rate)	Share of positive cases (infiltrated firms) correctly detected by the model.	$\frac{TP}{TP + FN}$
Specificity (True negative rate)	Share of negative cases (non-infiltrated firms) correctly detected by the model.	$\frac{TN}{TN + FP}$
Precision	Share of cases positively detected by the model which are actually positive.	$\frac{TP}{TP + FP}$

References

Barone G. and Mocetti S. (2014), "Natural Disasters, Growth and Institutions: A Tale of Two Earthquakes." *Journal of Urban Economics*, 84(C), pp. 52-66.

Bianchi P., Marra A., Masciandaro D., and Pecchiari N. (2020), OC and Firms' financial statements: Evidence from Criminal Investigation in Italy. *Bocconi Legal Studies Research Paper* No. 2017-59.

Brand, J. P. L. (1999), "Development, Implementation, and Evaluation of Multiple Imputation strategies for the Statistical Analysis of Incomplete Data Sets." *Ph.D. thesis*, Erasmus University.

Chengwei L., Yixiang C., Kazmi S.H.A. and Hao, F. (2015), Financial Fraud Detection Model Based on Random Forest. *International Journal of Economics and Finance*, vol. 7, no. 7,

Daniele, G. and Geys B. (2015), "OC, Institutions and Political Quality: Empirical Evidence from Italian Municipalities." *Economic Journal* 125(586), pp. 233-255.

Daniele V. and Marani. U. (2011), "OC, the quality of local institutions and FDI in Italy: A panel data analysis." *European Journal of Political Economy*, 27(1), pp. 132-142.

De Feo G. and De Luca G. (2013), "Mafia in the ballot box." *DEM Working Paper Series* No. 57.

De Simoni M. (2022), The financial profile of firms infiltrated by organised crime in Italy. *UIF, Quaderni dell'antiriciclaggio, Collana Analisi e Studi.*

Direzione Investigativa Antimafia (DIA) (2022), "Relazione del Ministro dell'Interno al Parlamento sull'attività svolta e sui risultati conseguiti dalla Direzione Investigativa Antimafia", luglio-dicembre 2022, available at: https://direzioneinvestigativaantimafia.interno.gov.it/relazioni-semestrali.

Donato L., Saporito A., and Scognamiglio A. (2013), "Aziende Sequestrate Alla Criminalità Organizzata: Le Relazioni Con Il Sistema Bancario." *Bank of Italy Occasional Paper* No. 202.

European Commission, Directorate-General for Migration and Home Affairs, Disley, E., Hulme, S., Blondes, E. (2021), "Mapping the risk of serious and organized crime infiltrating legitimate business : final report", *Publications Office,* available at https://data.europa.eu/doi/10.2837/64101

Fabrizi M., Malaspina P., and Parbonetti A. (2017), Caratteristiche e modalità di gestione delle aziende criminali. *Rivista di studi e ricerche sulla criminalità organizzata*, 3(1), pp. 47-66.

FATF (2018), "Concealment of Beneficial Ownership", *GAFI-Egmont eds.*

Haibo H. and Garcia E. A. (2009), Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, September 2009.

Kapoor S. and Narayanan A. (2023), "Leakage and the reproducibility crisis in machine-learning-based science" *Patterns*, vol. 4, no. 9, https://doi.org/10.1016/j.patter.2023.100804.

Jack M., Bosch Chen I. (2021), Impact of Organised Crime on the EU's Financial Interests, *Study*, European Parliament.

Lundberg L. (2017), "A Unified Approach to Interpreting Model Predictions", 31st *Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.

Maka K., Pazhanirajan S. and Mallapur S. (2020), Selection of most significant variables to detect fraud in financial statements. *Materials Today: Proceedings.*

Ministero dell'Interno (2023), "Le interdittive antimafia, 2021-2022", Dipartimento della pubblica sicurezza, Direzione centrale della polizia criminale, Servizio Analisi criminale, febbraio.

Mirenda L., Mocetti S., and Rizzica L. (2022), "The Economic Effects of Mafia: Firm Level Evidence." *American Economic Review*, 112 (8): 2748-73.

Mocetti S. and Rizzica L. (2021), "La criminalità organizzata in Italia: un'analisi economica", *Bank of Italy Occasional Paper* No. 661.

Peri G. (2004), "Socio-Cultural Variables and Economic Success: Evidence from Italian Provinces 1951-1991." B.E. *Journal of Macroeconomics*, 4(1), pp. 1-36.

Pinotti P. (2015), "The economic costs of OC: Evidence from Southern Italy." *Economic Journal* 125(586), pp. 203-232.

Ravenda D., Argilés-Bosch, J. M. and Valencia-Silva, M. M. (2015), Detection Model of Legally Registered Mafia Firms in Italy. *European Management Review*, 12: 23-39.

Sadgali, I., Sael N. and Benabbo F. (2019), Performance of machine learning techniques in the detection of financial frauds. *Procedia Computer Science.*

Shann H., Disley E. and Blondes E. L. (2019), Mapping the risk of serious and organised crime infiltrating legitimate businesses - Final Report, European Commission.

Sharma A. and Panigrahi P. (2013), A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *International Journal of Computer Applications* vol. 39, n.1.

Tianqi C. and Guestrin C. (2016), "XGBoost: A Scalable Tree Boosting System", *ACM*, http://dx.doi.org/10.1145/2939672.2939785

Transcrime (2015), Gli investimenti delle mafie. *Progetto PON sicurezza*, 2007-2013. Transcrime e Università Cattolica del Sacro Cuore.

Trevor H., Tibshirani R., and Friedman J. (2009), "The Elements of Statistical Learning", Springer New York, NY

UNODC (2011). Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes. *Research report*, United Nations Office on Drugs and Crime.

Van Buuren S. (2007). "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification." *Statistical Methods in Medical Research* 16:219–242.

Wyrobek J. (2020), Application of machine learning models and artificial intelligence to analyze annual financial statements to identify companies with unfair corporate culture. *Procedia Computer Science* 176, 3037–3046.

Zavitsanos E., Mavroeidis D., Bougiatiotis K., Spyropoulou E., Loukas L. and Paliouras G. (2022), "Financial misstatement detection: a realistic evaluation" *Proceedings of the Second ACM International Conference on AI in Finance (ICAIF '21),* Association for Computing Machinery, New York, NY, USA, Article 34, 1–9, https://doi.org/10.1145/3490354.3494453.