IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

A machine learning approach for the detection of firms infiltrated by organised crime in Italy[1]

Pasquale Cariello, Marco De Simoni and Stefano Iezzi,
Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

BANCA D'ITALIA
EUROSISTEMA

UIF
Unità di Informazione Finanziaria per l'Italia

A machine learning approach for the detection of firms infiltrated by Organised Crime in Italy

Pasquale Cariello, Marco De Simoni, Stefano Iezzi
Financial Intelligence Unit for Italy (UIF) – Bank of Italy

Rome, October 19, 2023

Disclaimer

The views expressed in this presentation are those of the presenter and do not necessarily correspond to those of the UIF or Banca d'Italia.

Outline

1. Research context and objectives
2. The dataset
3. Classification methodology
4. Main results
5. Robustness checks
6. Conclusions and future research

Research context

- Infiltrated firms are legally registered and seemingly legitimate businesses, controlled by organized crime (OC).
- Control is exerted through (Arlacchi, 2007):
  - ownership or management of the company by OC affiliates
  - use of financial resources derived from illegal activities
  - corrupt practices
  - and in some cases, even the adoption of illegal practices such as violence and intimidation.
- (De Simoni, 2022) finds that, in some cases, the use of corporate vehicles serves as a means to launder illegally obtained capital.
- Other studies (Ravenda *et al.*, 2015; Mirenda et al.,2022) consistently indicate that infiltrated firms display distinctive features in their financial statements. Insights from this stream of literature have stimulated the development of statistical models that try to detect infiltrated from non-infiltrated firms using financial reporting data.

Objective of the project

Capitalize on the findings of recent studies to build a machine learning algorithm to identify infiltrated businesses on the basis of financial statements and other information.

Two main innovations:

1. List of about 28,000 companies with a high probability of infiltration status:
   - combination of public sources and proprietary data from the UIF.
   - one of the most comprehensive censuses of companies controlled by OC in Italy.
   - previous studies typically use lists of infiltrated companies of smaller scale and with more uncertain links to OC.
2. A dataset of Italian companies (limited liabilities, joint stock companies, and other firms legally required to deposit financial statements) observed from 2010 to 2021 incorporating data on financial statements, debts to the financial sector, employment, and owners and administrators.

The algorithm provides a "risk" score for each capital company.

Overview of the panel dataset

Year	Infiltrated firms	Non-infiltrated firms	% of infiltrated firms
		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**

Variable selection

Drawing from the most relevant papers in this field,
we select a list of 32 variables/indicators that thoroughly characterize a firm's financial profile:

Economic Size (5 variables)
Total assets, revenues, equity, short-term liabilities, fixed assets

Equity and liquidity (7 indicators)
Cash/assets, short-term assets/short-term liabilities, cashflow, etc.

Indebtedness (4 indicators)
Granted loans over equity, Granted loans over revenues, Net debt over EBITDA, Total debt over assets

Discretionary elements (3 indicators)
Accrued liabilities over assets, Accrued incomes over assets, Inventory over assets

Investment and cost structure (5 indicators)
Tangibles/assets, Cost of rents/ revenues, Net purchases/revenues, Intermediate inputs/revenues, Capex

Employment (3 indicators)
Cost of labour over number of employees, Revenues over number of employees, Added value over number of employees

Profitability (5 indicators)
ROI, ROE, ROA, etc.

Three opacity indicators
- Ownership
- Management
- Miscellaneous attributes

Structural characteristics
Economic sector of activity (3-digit NACE code), Per capita province value added, Legal form.

Challenges in classification

- Some firms labelled as non-infiltrated may be connected to OC.
  - *Risk of potential bias in model training.*
    - ➢ *Use of the most up-to-date, reliable and comprehensive database to identify potentially infiltrated firms (UIF source; from now, for the sake of brevity, simply 'infiltrated')*
    - ➢ *The population of alleged legal firms is so large that this bias could be inconsequential.*
- Infiltrated firms in the target sample are disproportionately concentrated in Southern regions.
  - *Use of province-level dummies can improve the model's overall accuracy, but could result in an overly localized model and, therefore, less effective in classifying correctly firms in less affected provinces.*
    - ➢ *Use of a dual strategy: 1) per capita province-level value added; 2) no geographical covariate*
- High imbalance of the two classes (records of infiltrated firms 1.4%)
  - ➢ Stratified undersampling strategy with proportional allocation combined with clustering
  - ➢ Strata: combination of region and economic sector

Classification methodology - Data splitting

Total dataset
11,426,981 records (1,43% 'infiltrated')

Training set
80%

Test set
20% ~2,288k

Training set
60% ~210k

Validation set
20% ~53k

Data splitting is made with stratified under-sampling with clustering

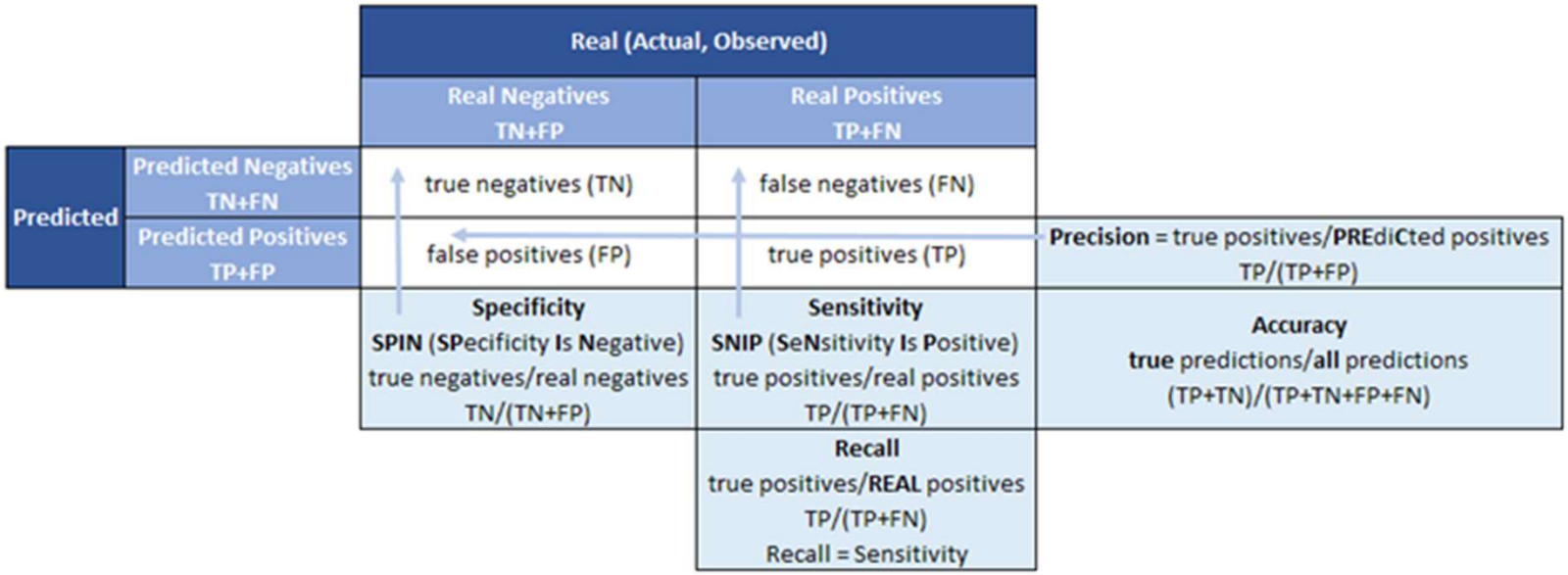
- Strata: infiltration status + region + economic sector
- Cluster: firm

Splitting is repeated 5 times to verify robustness of the results. We pick the 2nd best model.

Data Preparation & Model Calibration

- XGBoost (eXtreme Gradient Boosting) is the chosen algorithm to train the model;
- The dataset has been previously subjected to a very basic cleansing treatment in order to spot and resolve potential data inconsistencies;
- No imputation of missing data (only on Opacity variables), since Xgboost has the capability to effectively manage them;
- No transformation of the variables (only one-hot dummies on SECTOR);
- All monetary variables have been adjusted to 2021 constant prices;
- Use a randomized grids search with 50 random combinations to find the optimal combination of hyperparameters.

Some performance metrics



- Maximization of the performance measured by Recall (Sensitivity or TP rate)
- More tolerance on false positive (Specificity and Precision): it can be signal of potential infiltration in apparently legal firms.

Performance on unbalanced test set

Comparison with 'baseline' models;
(DummyClassifiers makes predictions based only on target distribution, ignoring the input features).

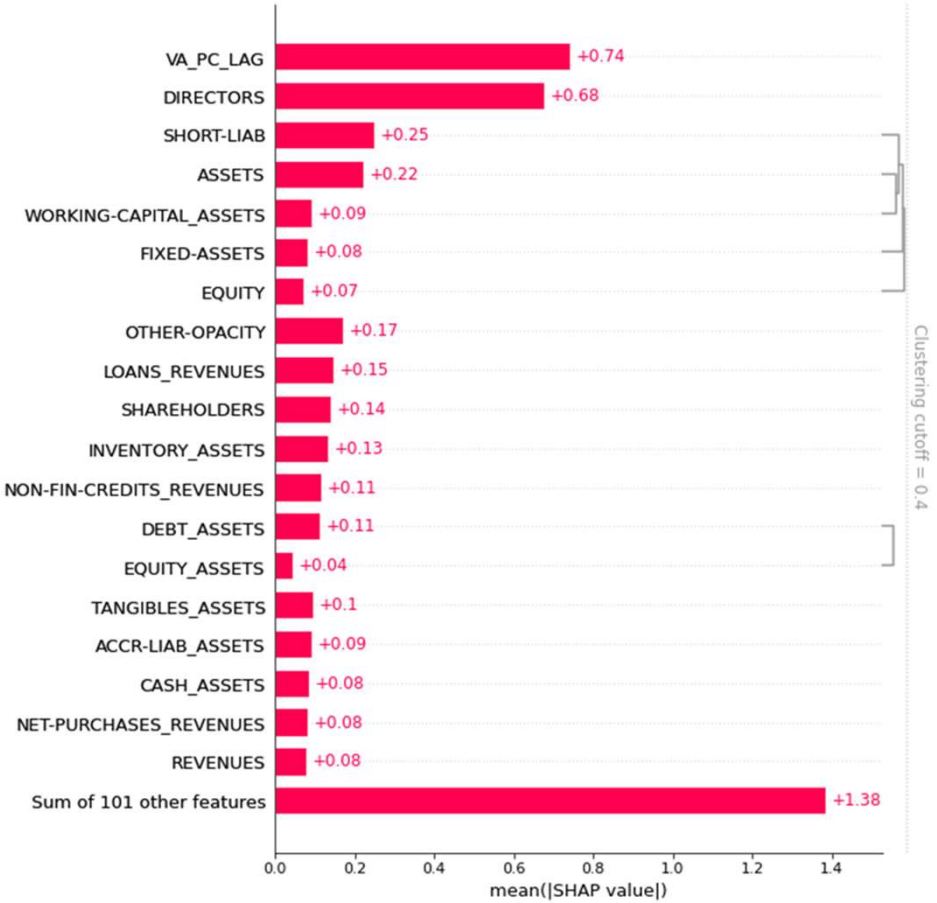
TEST SET	Accuracy	Sensitivity (Recall)	Specificity	PR-AUC	F2
XGB	0.743	0.756	0.742	0.678	0.752
Stratified	0.501	0.490	0.511	0.493	0.491
Most frequent (0)	0.986	0.000	1.000	0.492	0.000
Uniform	0.500	0.499	0.500	0.497	0.494
Less Frequent (1)	0.014	1.000	0.000	0.492	0.829

Risk score computation

Frequency distribution of estimated risk score
Year 2021

Risk score	N	%
Up to 0.50	729,433	78.3
From 0.51 to 0.80	129,799	13.9
From 0.81 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

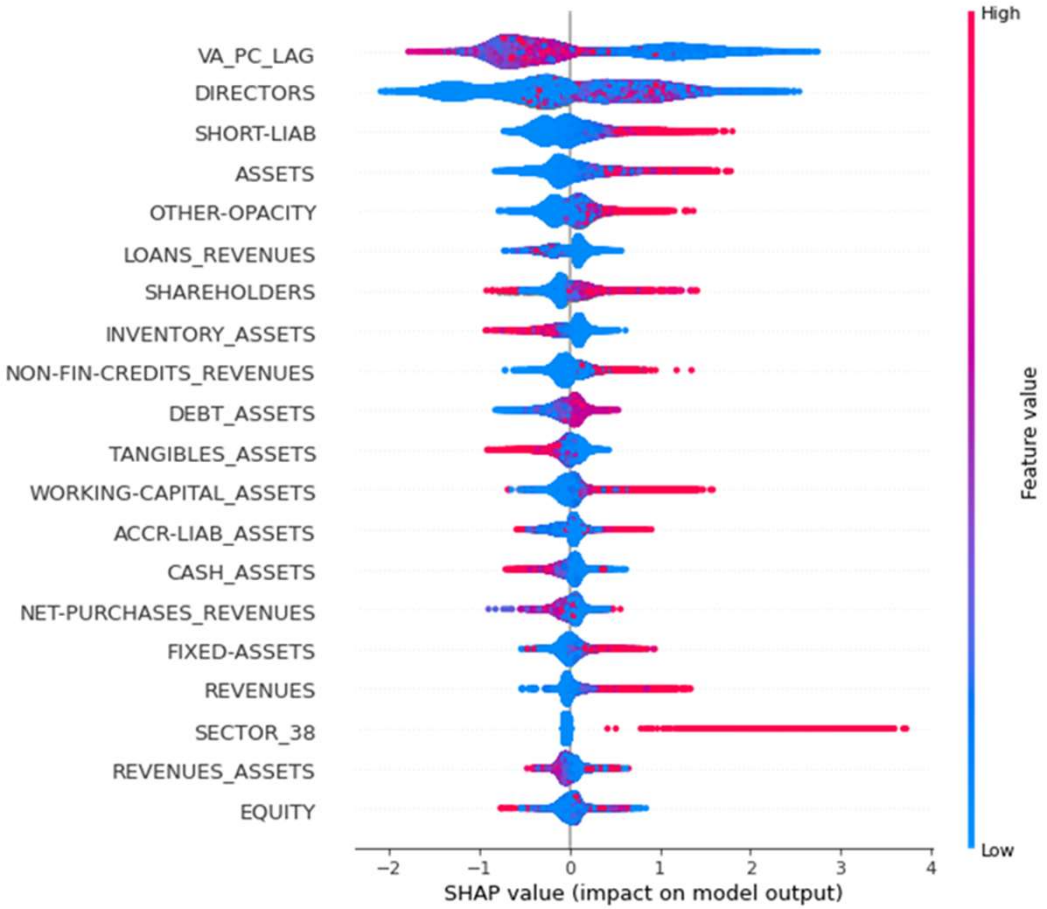
Variable importance - SHAP



Most influencial groups:

- Geo & Sector
- Opacity
- Size (Assets and Revenues)
- Equity and Liquidity

Variable importance - SHAP (2)



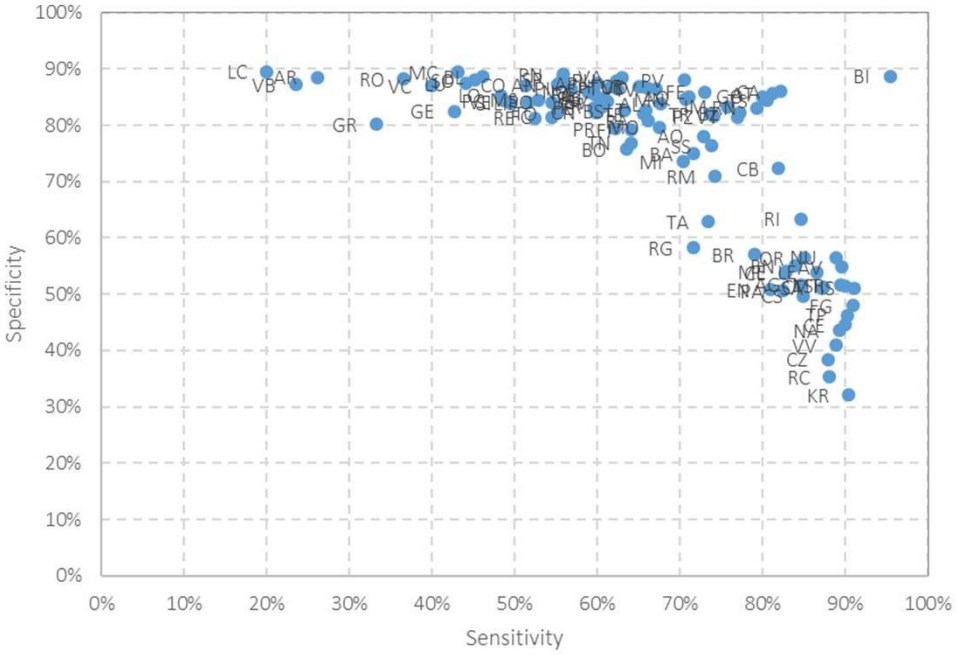
- As the "size" (assets, revenues, fixed assets, and short liabilities) increases, the probability of predicting infiltration also rises.
- Investment indicators (specifically, tangible assets over total assets) and liquidity indicators (namely, cash over total assets) have a negative effect on the likelihood of model prediction.
- Research has highlighted that companies infiltrated by organized crime are, on average, larger, more indebted and less liquid (Bianchi et al., 2022).

Test set performance by geographical breakdown

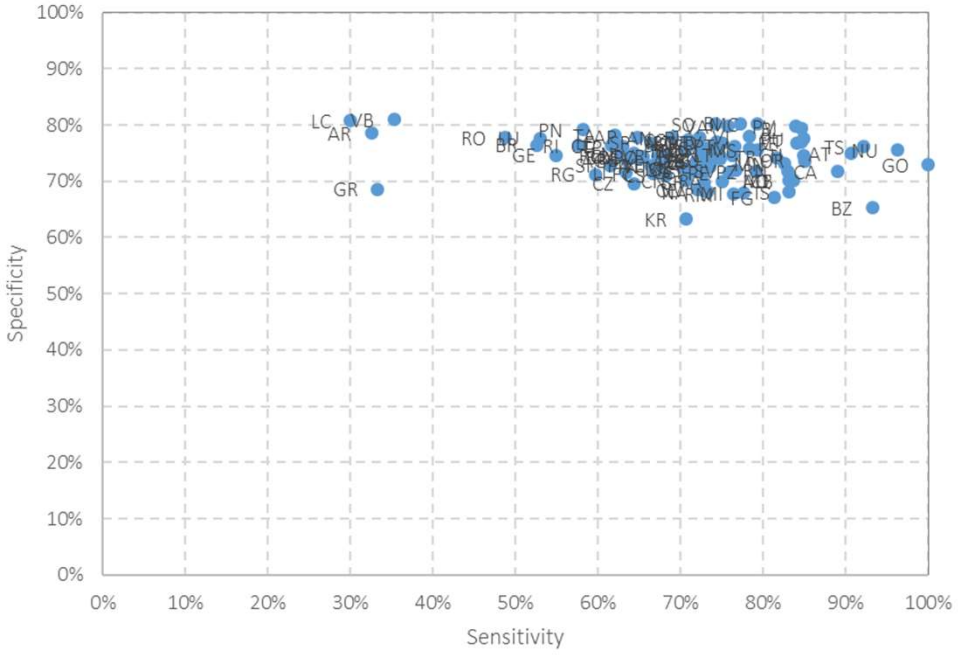
	With per-capita province-level value added		Without any geographical variable	
	Recall	**Specificity**	**Recall**	**Specificity**
North-West	0.650	0.805	0.737	0.727
North-East	0.624	0.830	0.713	0.759
Centre	0.686	0.782	0.742	0.707
South and Islands	0.858	0.582	0.711	0.692
Total	**0.755**	**0.742**	**0.723**	**0.719**

Performance on test set by province

Per capita province-level value added used as a covariate



No geographical information used as a covariate



Conclusions and future development

- We developed of a Machine Learning algorithm to detect legally registered firms *potentially* infiltrated by OC (risk score)
- Sample of about 28 thousand Italian firms considered to be infiltrated with high probability (Primary source is the UIF archives)
- Highly varied list of financial and budget indicators and variables, identified on the basis of the latest literature on criminal infiltration in real economy.
- Use of information on ownership and managers to build 3 opacity indicators.
- Infiltrated firms compared with stratified random samples of alleged legal firms in order to train and test the model.
- Model performance on test set: Recall 75.6%, Specificity 74.2%.
- Risk score resulting from the algorithm as an *additional* red flag indicator for UIF institutional functions.
- Potential scope for sharing the risk score information with other selected actors involved in AML and/or the fight against economic crime.
- Opportunities for future research: alternative methods to manage unbalancing; adopt a dynamic approach; variable transformations to cancel out geographic/sectoral effects.

References

- Arlacchi P. (2007), La Mafia imprenditrice, Il Saggiatore.
- 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.
- Mirenda, Litterio, Sauro Mocetti, and Lucia Rizzica (2022), "The Economic Effects of Mafia: Firm Level Evidence." *American Economic Review*, 112 (8): 2748-73.
- De Simoni M. (2022), The financial profile of firms infiltrated by organised crime in Italy. *UIF, Quaderni dell'antiriciclaggio, Collana Analisi e Studi*.

Thank You
For Your Attention