DNB Data Science Conference

"Central bankers go data driven: applications of AI and ML for policy and prudential supervision"

A Machine Learning approach for the detection of firms infiltrated by organised crime in Italy

BY P. CARIELLO, M. DE SIMONI AND S. IEZZI – UNITA' DI INFORMAZIONE FINANZIARIA (UIF) – BANCA D'ITALIA

Amsterdam, 12 May 2022

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

Outline

- Background
- ML Pipeline
- Data
- Model
- Preliminary results
- Possible applications for AML and prudential supervision

Background

- Estimates by the United Nations Office on Drugs and Crime show that in 2009 organized crime’s (OC) revenues amounted to 3.6% of the world’s GDP (UNODC, 2011).

- European Council, in 2019 criminal revenues in the main criminal markets amounted to 1% of the EU's GDP, i.e. €139 billion

- the proceeds from mafia groups’ illegal activities could represent up to 2 per cent of national GDP, as shown in a study by Transcrime, in cooperation with the Italian Ministry of the Interior (Transcrime, 2015)

→ Some studies find that infiltrated firms, from a financial point of view, exhibit a peculiar financial statement’s structure, at least with regard to some of its dimensions.

→ These findings have given rise to the development of statistical models aiming to discriminate between infiltrated and non-infiltrated firms on the basis of financial reports and, ultimately, to detect apparently lawful firms, which are actually controlled by organized crime.

Our contributions

1. we build a large firm-level dataset for Italy spanning from 2010 to 2020 by merging financial statement information collected from multiple sources (~3,2 mln of records). This highly varied source of data allows us to construct a large set of financial variables and indicators;

2. we use a unique sample of about 1,800 firms that are infiltrated with a high degree of confidence, which make our study substantially more robust than the existing research on this topic;

3. we resort to a machine learning approach with the aim of building a classifier capable to identify legally registered firms potentially infiltrated by organized crime with superior performances.

ML Pipeline

Data Preparation

- Data preparation (collect, integrate, clean, impute)
- Variables selection
- Down-sampling
- Splitting

Train & compare models

- Train models
- Hyper-parameter-tuning (with cross-validation)
- Model testing
- Model comparison
- Model use

The data

Construction of the sample of infiltrated firms

Total firms: 1,786
Total records: 6,294

List C

- 1,223 firms
- (investigated administrators/stakeholders)

List B

- 340 firms
- (archive of seized businesses)

List A

- 223 firms
- *(investigated by Italian law enforcement unit specialised in the fight against OC and terrorism)*

List A & B - we only use data for all years up to the second before the seizure
List C - we discard all data from previous years where colluded stakeholders or administrators take control of the firm

The data/2

Missing data treatment

For **alleged legal firms**, since we have a very large number of records, we decide to use a complete case analysis (CCA) approach by removing missing data using listwise deletion, i.e. deleting data for all cases that have missing data for any variable. We end up with a sample of only 33.5 per cent of complete records and 44 per cent of firms.
The full sample is not much different from the sample of complete cases according to the distribution of firms by sector and region.

For **infiltrated firms**, since we have a limited sample, we only remove records with missing data for the province and sector categorical variables or with more than 6 missing items. Then we apply a fully conditional specification (FCS) method to impute the remaining missing values. This selection reduces the number of records from 9,294 to 6,294 (the number of firms drops from 2,293 to 1,786), with no impact on the distribution by sector and region.

The data/3

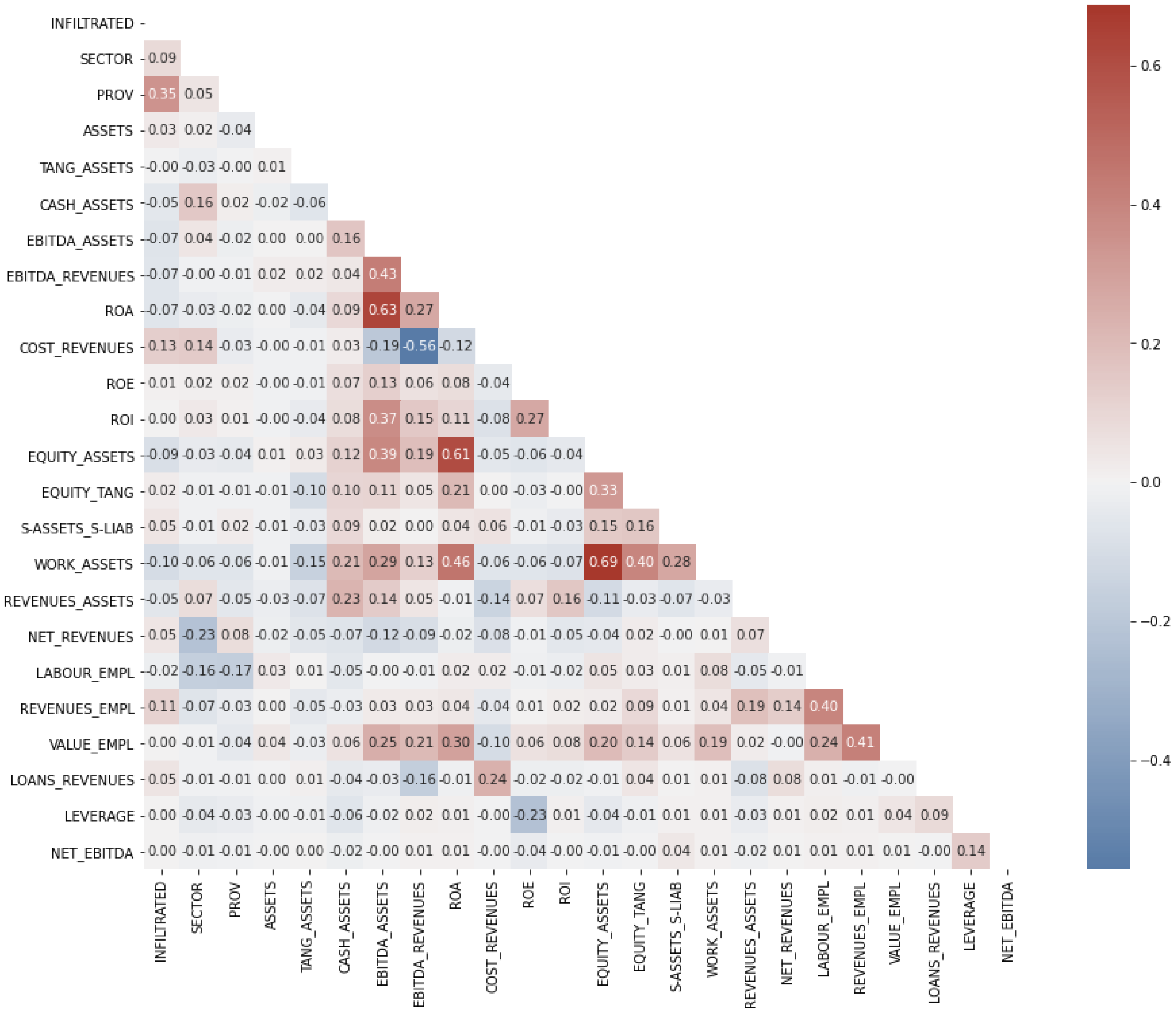
		Cardinality
Infiltrated firms	Annual financial stamements	6.294
	Firms	1.786
	Statements per firm *(mean)*	3,5
Non-infiltrated firms	Annual financial stamements	3.224.204
	Firms	746.843
	Statements per firm *(mean)*	4,3

Variables selection

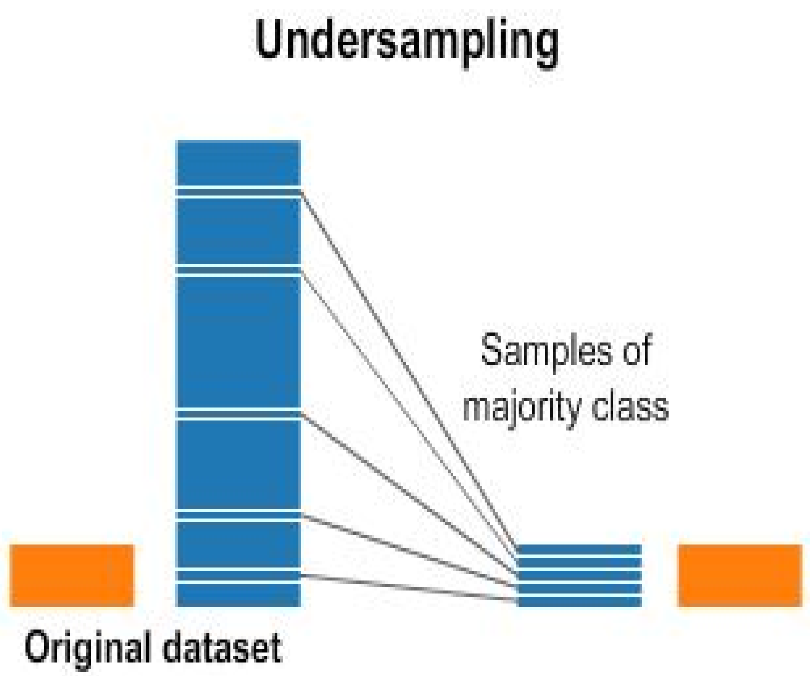
Dimension of analysis	Variable	Source	Abbreviation
Sector of activity	3-digit NACE code	Central business registry / National Statistical Institute	SECTOR
Location	Province of location		PROV
Size	Assets	Central business registry	ASSETS
	Revenues		REVENUES
	Equity		EQUITY
	Tangibles		TANGIBLES
	Short term liabilities		SHORT_LIAB
Equity and liquidity ratios	Cash over assets	Central business registry	CASH_ASSETS
	Equity over assets		EQUITY_ASSETS
	Equity over tangibles		EQUITY_TANG
	Short-term assets over short-term liabilities		S-ASSETS_S-LIAB
	Revenues over assets		REVENUES_ASSETS
	Working capital over assets		WORK_ASSETS
Indebtedness	Leverage (granted loans over equity)	Central business registry / Central Credit Registry	LEVERAGE
	Granted loans over revenues		LOANS_REVENUES
	Net debt (granted loans - cash) over EBITDA		NET_EBITDA
Profitability	EBITDA over revenues	Central business registry	EBITDA_REVENUES
	EBITDA over assets		EBITDA_ASSETS
	ROI		ROI
	ROE		ROE
	ROA		ROA
Investment (internal vs external resources) and cost structure	Tangibles over assets	Central business registry	TANG_ASSETS
	Cost of rents and leases over revenues		COST_REVENUES
	Net purchases over revenues		NET_REVENUES
Employment	Cost of labour over number of employees	Central business registry / National Institute for Social Security database	LABOUR_EMPL
	Revenues over number of employees		REVENUES_EMPL
	Added value over number of employees		VALUE_EMPL

Variables selection

Pairwise linear correlation of the variables of the model



Managing unbalanced sample



» High imbalance between records for infiltrated and non-infiltrated firms: ~1/500 ratio!
→ low ability to recall infiltrated firms (sensitivity)

» We use an *under-sampling* approach:
- ✓ Reduction of non-infiltrated firms by sampling;
- ✓ Strata are defined according to the combination of year, region and sector of activity of the firm;
- ✓ We choose a proportion of infiltrated firms of about 40% of the total.

Models comparison

Preliminary results

Model	**XGBoost**	**Random forest**	**Logistic**	**Neural Network**
Accuracy	0.86	0.82	0.7	0.73
Precision (sensibility)	0.84	0.80	0.66	0.58
Recall (sensitivity)	0.81	0.75	0.53	0.69
F1-score	0.83	0.77	0.59	0.63

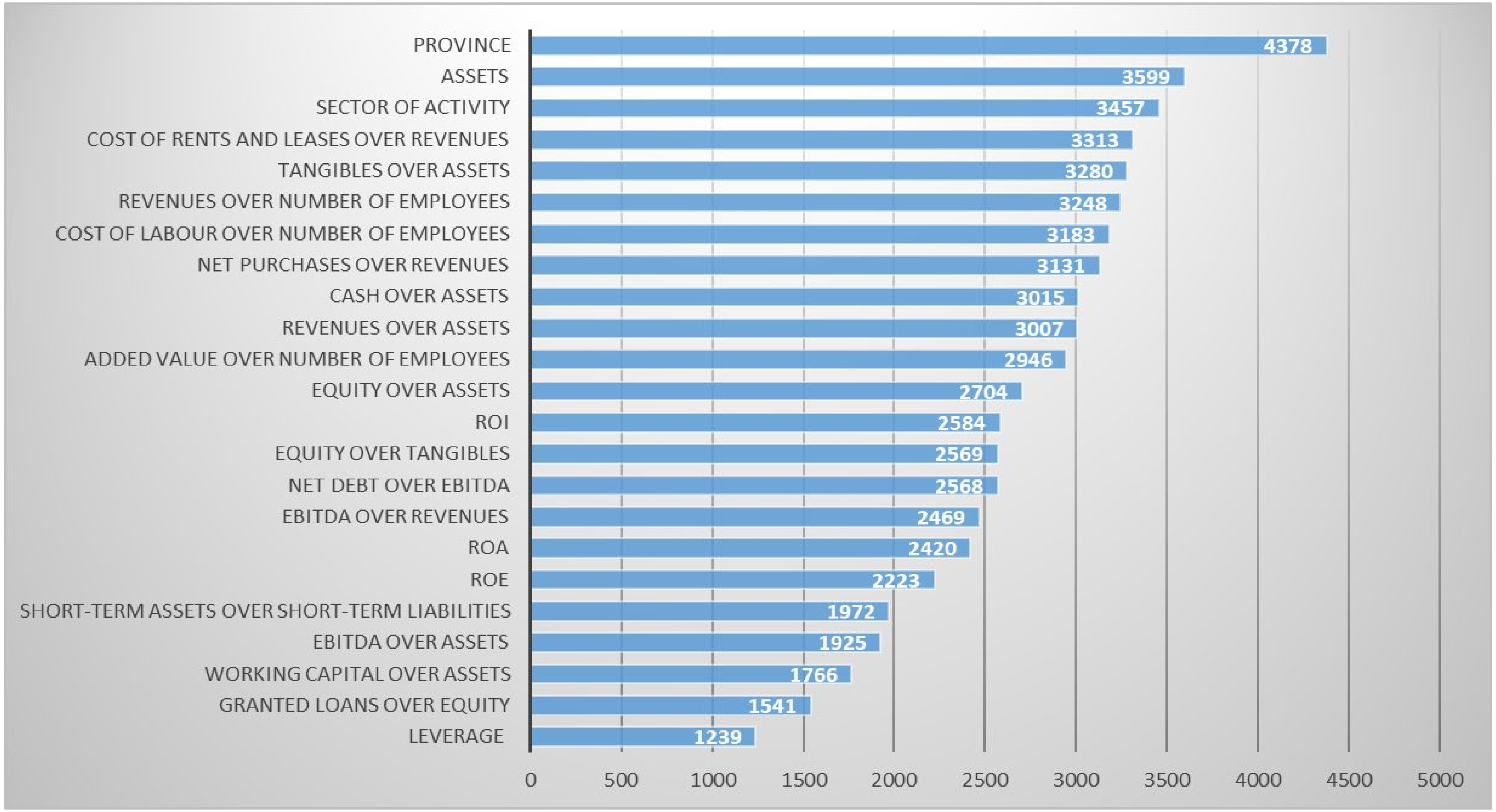
XGBoost Performaces

Preliminary results

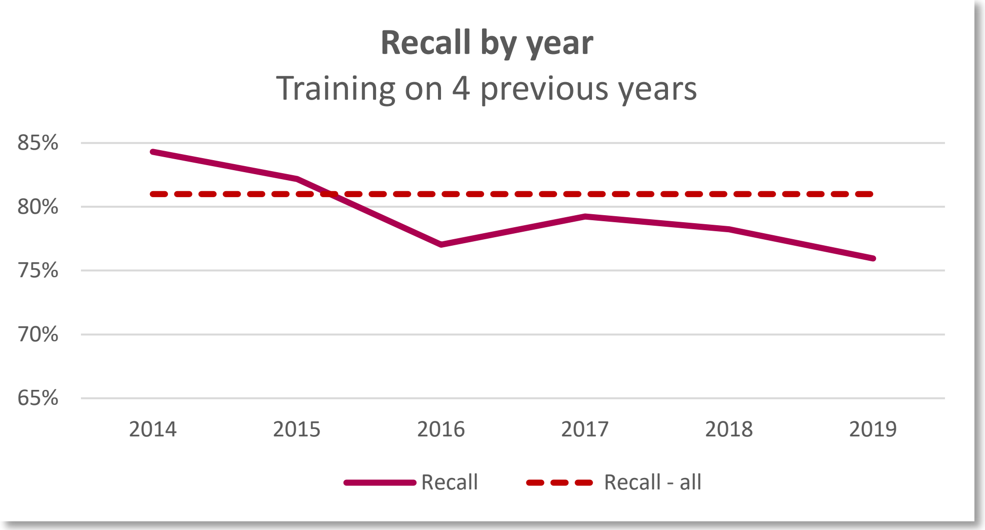
Accuracy score is: 0.865
Precision score is: 0.843
Recall score is: 0.813
F1 score: 0.828

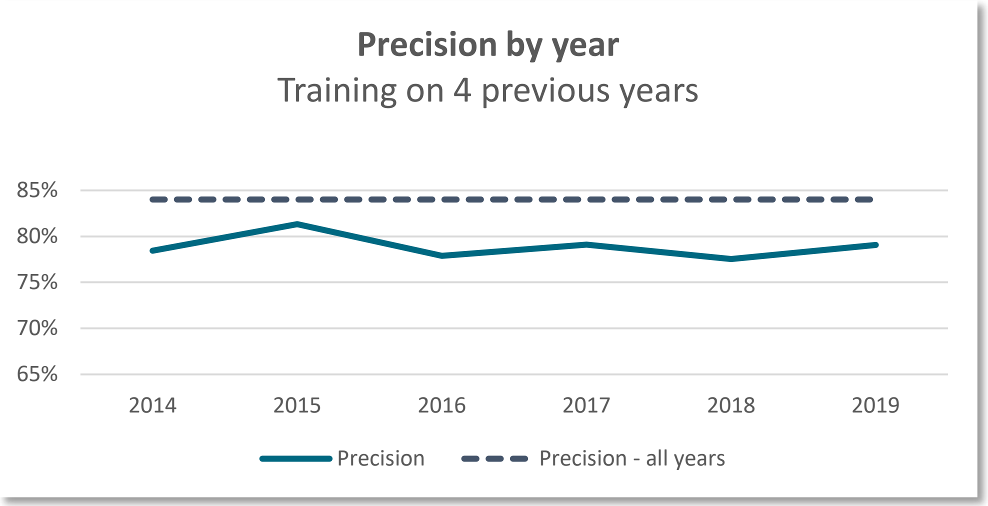
- n_estimators: 500 (specifies the number of decision trees to be boosted)
- max_depth=10 (it limits how deep each tree can grow).
- learning_rate=0.1: (it is a regularization parameter that shrinks feature weights in each boosting step)
- Other parameters are left to default values

Relative importance of features



‘Stability’ test over years





Applications

Computation of the risk score 472,539 firms, for which we have complete records in the most recent years:

Table 6. Frequency distribution of estimated risk score – years 2018-2020

Risk score	N	%
Up to 0.5	423,360	89.6
From 0.5 to 0.8	21,983	4.7
From 0.8 to 0.95	14,571	3.1
From 0.95 to 0.99	7,797	1.7
Over 0.99	4,828	1.0
Total	**472,539**	**100.0**

Possible uses:

1) to prioritise work within the central AML authority, as it may signal a potential involvement of high-risk companies in the financial conducts that are reported as suspicious by AML obliged entities.
2) To compute an aggregate risk indicator both at a geographical or sectoral level, which may provide interesting insights, for instance, within the National Money Laundering Risk Assessment.
3) To derive the financial exposure of each banking institution towards risky companies.

Further developments

1) Expanding the sample of infiltrated firms, by resorting to other sources;
2) Adding further financial and non-financial information gathered from other sources to the set of explanatory variables;
3) Using alternative sampling methods for imbalanced learning, like SMOTE, ADASYN or ensemble learning techniques;
4) Adopting multiple imputation techniques for alleged legal firms, in order to compute the risk score for a larger portion of Italian registered limited liability companies, thus widening the scope of application for AML and prudential supervision purposes.

Thank You

For Your Attention