Joint workshop with experts from EIOPA, the ECB, NCAs, NCBs
and industry representatives on Insurance Reporting
2024-10-24

BANCA D'ITALIA

A supervised record linkage approach for anomaly detection in insurance assets granular data

Vittoria La Serra
Directorate General for economics, statistics and research
Statistical data collection and processing Directorate

In 2021: a preliminary version of this work was presented

In 2024: an enriched version, published in a scientific journal
(La Serra, V., Svezia, E. (2024) «Quality and Quantity»)

BANCA D'ITALIA

- European Insurance Corporations (ICs) quarterly report to national supervisory authorities (NCAs) and central banks (NCBs) since 2016 (Solvency II Directive).
- Asset-by-asset information is reported by the ICs in a specific "list of assets" template (S.06.02) which is used for statistical purposes by central banks and for supervisory purposes by competent authorities.
- The template can be enriched with different external sources → CSDB, RIAD, other sources internally used at NCA/NCB level.

A database of reported assets, where each asset of an IC has:

- **An identification (ID) code** → required to be kept stable and consistent over time
- **A set of qualitative and quantitative features** → issuer/counterpart sector, issuer/counterpart area, CIC, reporting sector of IC, original currency, asset's market value…

- **The issue in the data:** errors in the reporting of ID codes might occur.

  We can observe changes in the ID codes of the reported assets, from quarter to quarter, even if, by regulations, these should remain unique and consistent over time.

- **Consequences:**
  - two assets from two subsequent quarters are perceived as different when in reality they are the same;
  - we lose the information on the past history of the asset;
  - decrease in quality of IC statistics to be compiled and disseminated.

- **The goal:** to build a model that is capable of identifying pairs of assets that do not share the same ID code but actually refer to the same asset.

In the Italian database:

- On average: 70,000 assets reported at each quarter;
- On average, in each quarter, there is a turnover of 8% for the reported assets:
  - New purchased assets
  - Sold assets
  - Reporting errors in ID codes

8% is taken as the maximum expected percentage of cases of anomaly: a limited impact on data quality, but still non-negligible.

A record linkage approach

- Select two datasets containing assets from two subsequent quarters Q_t and Q_{t+1}.
- Build a comparison matrix to compare all pairs of assets reported by the same IC on observed features (qualitative/quantitative) via distance measures.

Asset codes		Target	Distance measures on the observed features		
Quarter Q_t	Quarter Q_{t+1}	Status	$dist_1$	…	$dist_k$
Code A	Code A	Match	…	…	…
Code A	Code B	Non-match	…	…	…
Code B	Code B	Match	…	…	…
…	…	…	…	…	…

Idea: same assets are similar on the observed features!

A record linkage approach

Each couple of assets can either be a:

- Match (the two assets share the same ID code)
- Non-match (the two assets have different ID codes)

Asset codes		Target	Distance measures on the observed features		
Quarter Q_t	Quarter Q_{t+1}	Status	$dist_1$	…	$dist_k$
Code A	Code A	Match	…	…	…
Code A	Code B	Non-match	…	…	…
Code B	Code B	Match	…	…	…
…	…	…	…	…	…

Fit supervised classification models on the matrix:

- The **target** to predict: the binary status variable *{match, non-match}*
- The **covariates**: the computed distance measures

Asset codes		**Target**	**Distance measures on the observed features**		
Quarter Q_t	**Quarter Q_{t+1}**	**Status**	$dist_1$	…	$dist_k$
Code A	Code A	Match	…	…	…
Code A	Code B	Non-match	…	…	…
Code B	Code B	Match	…	…	…
…	…	…	…	…	…

BANCA D'ITALIA

Fit supervised classification models on the matrix:

- A model will predict the probability that each couple of records is a *match*.
- Fixing a probability threshold, we can predict the target & assess the model's performance through: true positive rate, false positive rate, etc.

Asset codes		Target	Predicted target	Outcome
Quarter Q_t	Quarter Q_{t+1}	Status		
Code A	Code A	Match	Match	Correct prediction
Code A	Code B	Non-match	Non-match	Correct prediction
Code B	Code B	Match	Non-match	Prediction error!
…	…	…	…	…

- From the Italian database, 13 couples of subsequent quarters are considered (for a timespan of more than three years).
- For each couple $(t, t+1)$, a comparison matrix is built to compare all assets from t to all assets from $t+1$, referring to the same IC (around 150mln of couples for each couple of quarters).

- **Three models are tested**: logit, random forest and neural network. Tests are run on datasets for all couples of quarters and on differently sampled datasets, to ensure robustness.
- Models' performance show the superiority of the **random forest** over the others.

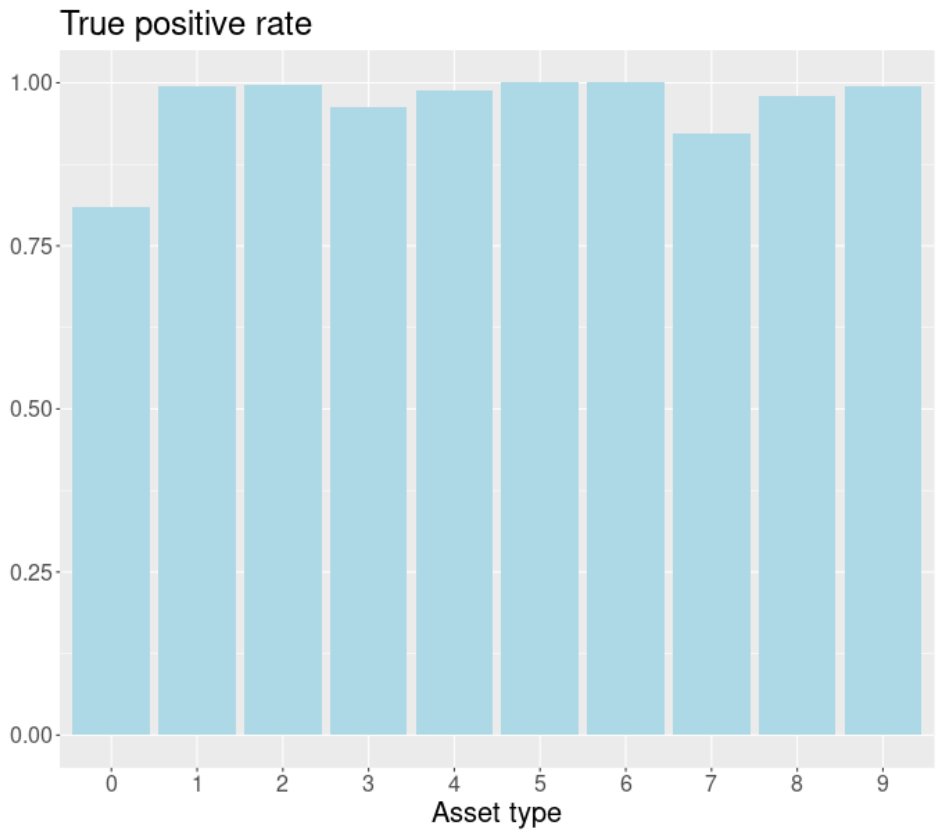
The random forest: **general results***

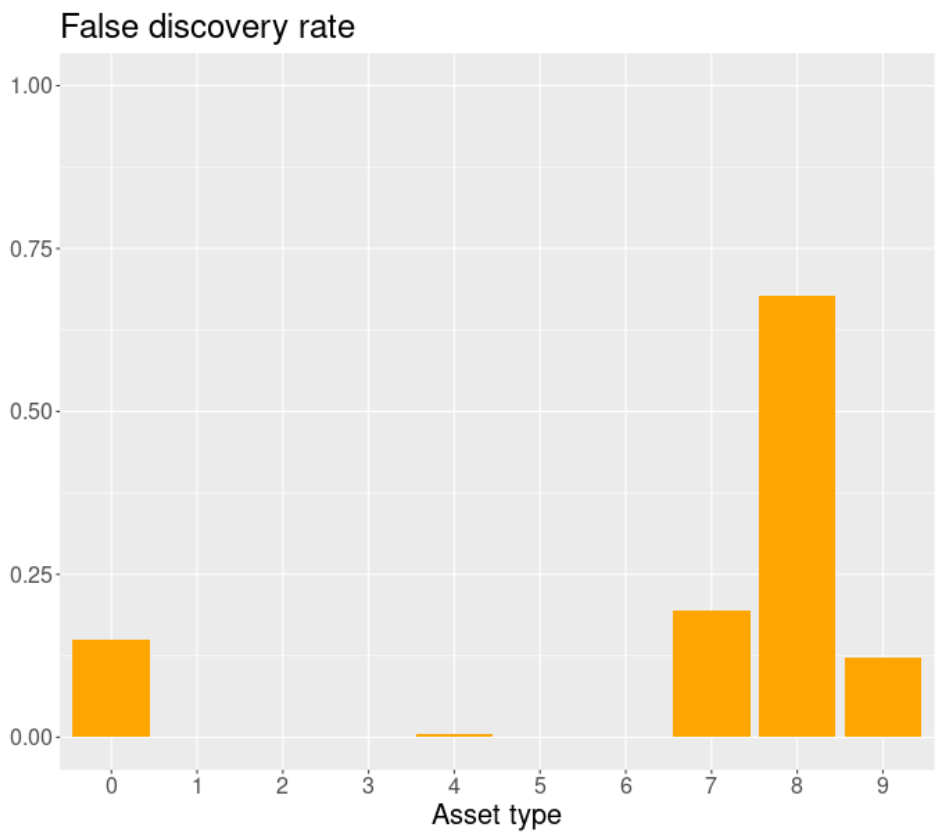
Metric	Description	Observed value
Accuracy	A general model's performance	99.5%
Balanced accuracy	A general model's performance, more robust to unbalances in the data	99.0%
True positive rate	The rate of correctly detected errors	98.6%
False discovery rate	The rate of non-errors being classified as errors: a cost measure for the model	8.9%

*for a 5%-95% unbalanced dataset on the target variable and a fine-tuned probability threshold of 0.3

The random forest: **asset type (CIC) detailed results**

The asset type (third character in the CIC)	
1	Government bonds
2	Corporate bonds
3	Equity
4	Investment funds Collective Investment Undertakings
5	Structured notes
6	Collateralised securities
7	Cash and deposits
8	Mortgages and loans
9	Property
0	Other investments





The random forest: **asset type (CIC) detailed results**

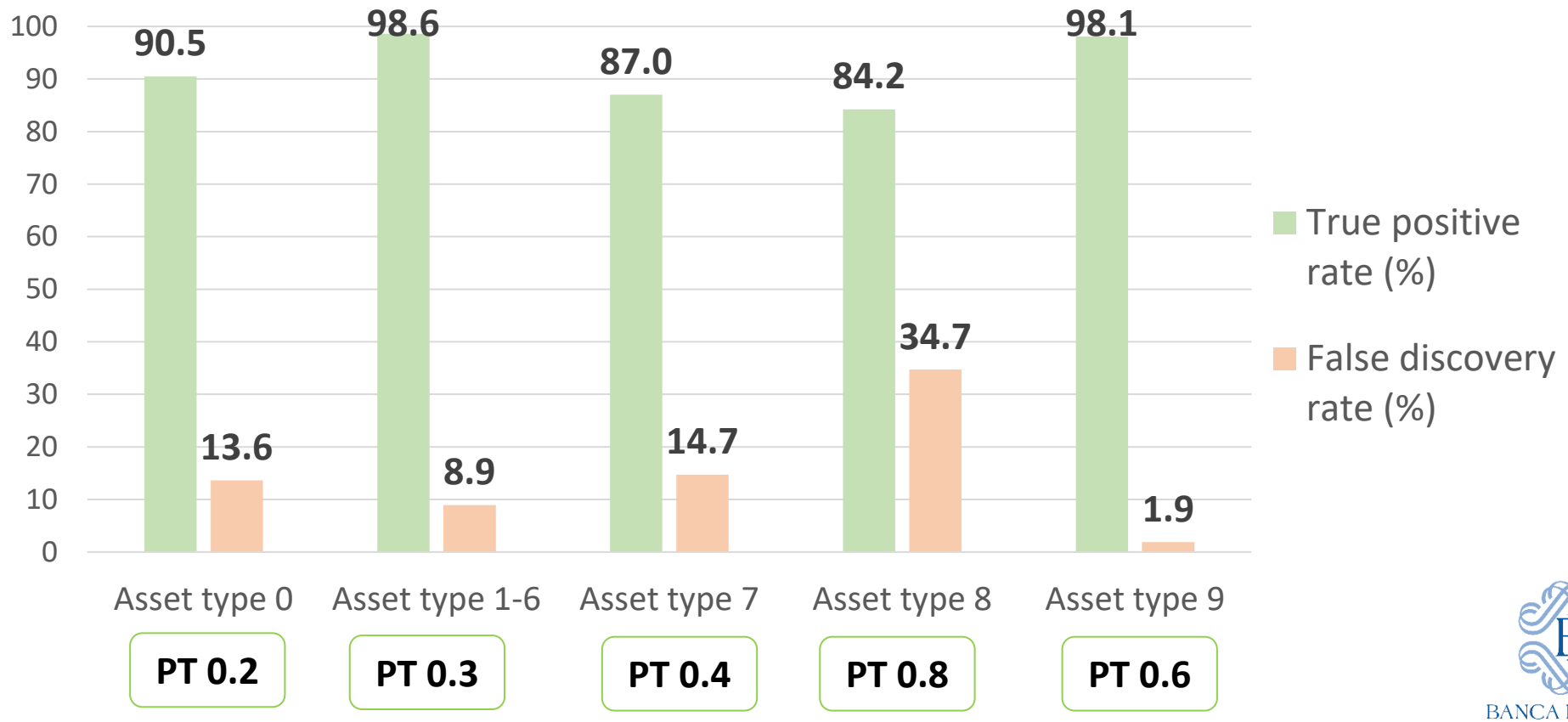
Possible reasons for such variability in the performance:

- the **different volume** of data in each asset type class in the database, being small for non-securities;
- the presence or absence of **standard for the identification code**: ISIN codes are widely available for securities and reporting by standard is strongly recommended by regulators; IDs for assets that are not securities are often chosen arbitrarily by reporting ICs;
- available features for non-securities assets are **less informative** and many features used in the training matrix are securities-specific (e.g. issue date, issuer sector, issuer area).

The random forest: **asset type (CIC) detailed results**

The asset type (third character in the CIC)	
1	Government bonds
2	Corporate bonds
3	Equity
4	Investment funds Collective Investment Undertakings
5	Structured notes
6	Collateralised securities
7	Cash and deposits
8	Mortgages and loans
9	Property
0	Other investments

Fine-tuned probability threshold (PT) for each asset type



BANCA D'ITALIA

Using the model in production

- The actual performance of the proposed methodology must be validated through the **crosscheck with the ICs** of the estimated errors in a quarter during a real data production round.
- Interaction with the Italian NCA helps us exclude some cases of false discovery, before reaching out to the ICs. This **deterministic filtering** increasingly improves our model's effectiveness & lightens the burden on the ICs.

- The model was used during real production rounds and, after the first quarters of application, it demonstrated its effectiveness: **60.2%** of the detected cases was indeed anomalous; **2.3%** was made of false discoveries; the remaining **37.5%** is still under investigation.

Conclusions

- **Data Quality Management** (DQM) processes in central banks are necessary to ensure high quality in the disseminated statistics and machine learning models are emerging to approach such issues.
- An **automated method** to detect the presented issue in IC data is necessary to ensure high quality of insurance statistics, given the volume of increasingly granular databases and the impact that errors have on compiled and disseminated statistics.
- The proposed methodology returns **good results** to reach the goal with **high performance**; periodical maintenance is required: monitoring of the model's performance and updated training on more recent data if performance decreases.
- **Interaction among data stakeholders can be intensive in some cases**, due to the different perspective on data: **supervisory vs statistical focus. Cross-checking** with the ICs during real data production rounds can be **time and resource consuming**.

BANCA D'ITALIA

Thank you for your attention.

Vittoria La Serra – Vittoria.laserra@bancaditalia.it

Assessing the performance of supervised binary classification models		
Performance metric	Description	Formula
True Positive Rate (TPR)	Correctly classified cases of match	TP/(FN+TP)
True negative rate (TNR)	Correctly classified cases of non-match	TN/(TN+FP)
Accuracy	Correctly classified cases	(TN+TP)/N
Balanced accuracy	A general model performance, more robust to unbalance in the data	(TPR+TNR)/2

Confusion matrix for a given probability threshold for classification		
	Actual non-match	**Actual match**
Predicted non-match	TN	FN
Predicted match	FP	TP

BANCA D'ITALIA

Assessing the performance of supervised binary classification models		
Performance metric	Description	Formula
False positive rate (FPR)	The erroneously classified cases of match: the missed non-matches	FP/(FP+TN)
False discovery rate (FDR)	The erroneously classified cases of match: a cost measure for the model	FP/(FP+TP)
ROC curve	A general model performance	FPR and TPR varying with the threshold for classification
AUC index	The area under the ROC curve	

Confusion matrix for a given probability threshold for classification		
	Actual non-match	**Actual match**
Predicted non-match	TN	FN
Predicted match	FP	TP

BANCA D'ITALIA