BANCA D'ITALIA

Statistical matching for anomaly detection in insurance assets granular reporting

Vittoria La Serra, Emiliano Svezia

Bank of Italy

Statistical Data Collection and Processing Directorate

Basel, 26th August 2022

- Data Quality Management (DQM) in central banks: necessary to ensure high quality in disseminated statistics.
- Automation of DQM processes is crucial:
  - to manage the volume of increasingly granular databases
  - to ensure resilience in situations of human resources constraints (pandemic)
- Machine Learning (ML) models: emerging to solve DQM issues

- **Proposal**: a record linkage approach using ML models to deal with a DQM issue on insurance granular assets data.

- European Insurance Corporations (ICs) quarterly report to national supervisory authorities and central banks since 2016 (Solvency II Directive).

- Asset-by-asset information is provided in template S.06.02 and used for statistical purposes by central banks.

- Each asset of an IC is reported with:
  - An identification (ID) code → required to be kept stable and consistent over time
  - A set of qualitative and quantitative features

- The DQM issue: reporting errors in ID codes might occur.

- Consequences:
  - two assets from two subsequent quarters are perceived as different when in reality are the same;
  - 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.

A record linkage approach

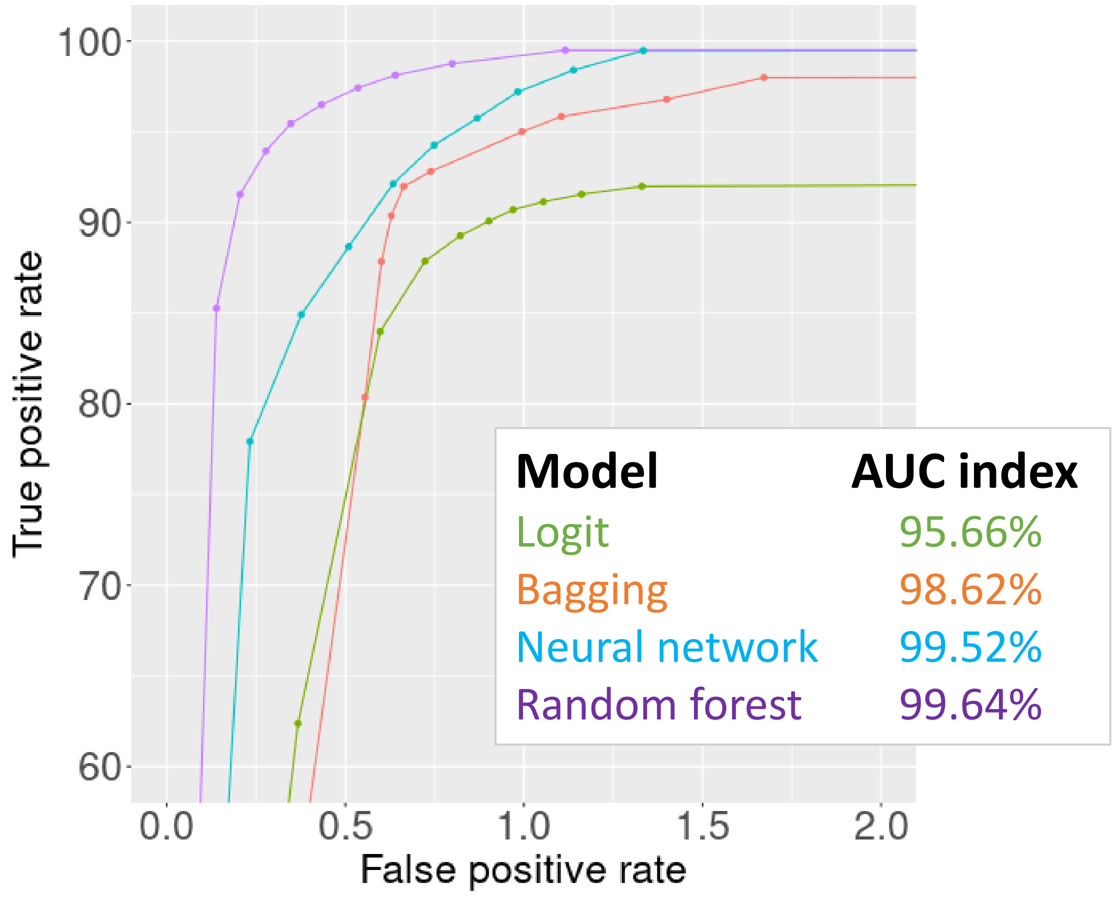
- Select two datasets containing assets from two subsequent quarters Q_t and Q_{t+1}.
- Same assets are similar on the observed features → build a comparison matrix to compare all pairs of assets reported by the same IC on observed features (qualitative/quantitative) via distance measures.
- Fit supervised classification models on the matrix, where the target variable to predict is the binary status of each pair: *{match, non-match}*.

Asset codes		**Target**	**Distance measures on the observed features**			
Quarter Q_t	**Quarter Q_{t+1}**	**Status**	**Nominal**	**Ordinal**	**Numerical**	**Textual**
Code A	Code A	Match	…	…	…	…
Code A	Code B	Non-match	…	…	…	…
Code B	Code B	Match	…	…	…	…
…	…	…	…	…	…	…

- From the Italian database, assets from two subsequent quarters are selected and compared, building the comparison matrix.
- 70,000 assets reported on average at each quarter → billions of pairs of assets to compare.

- Different supervised classification models have been trained and tested:
  - Logit (benchmark), bagging, random forest, neural networks.
  - Fine tuned for different hyperparameters combinations (e.g. number of trees, number of hidden layers).
  - Repeatedly fitted on differently unbalanced datasets (w.r.t. the target) to ensure robust results.

AVERAGE* ROC CURVES



Model	**AUC index**
Logit	95.66%
Bagging	98.62%
Neural network	99.52%
Random forest	99.64%

*Different percentages of unbalance

TEST RESULTS
for the Random Forest

- Hypothesizing 5-95% unbalance in the target
- Selecting a probability threshold of 0.2

Balanced accuracy	99% ↑
Correctly classified cases of match (True positive rate)	98.5% ↑
Erroneously classified cases of non-match (False discovery rate)	9.5% ↓

Conclusions

- The proposed methodology returns promising results to reach the goal with high performance.
- An automated method to detect errors in reported ID codes is necessary to ensure high quality of insurance statistics, given: the need for resilience in DQM processes; the volume of IC assets; the impact that such errors have on compiled statistics.

Further developments

- Improvement in the model training phase: considering all the available Italian data since 2016, not only focusing on two subsequent quarters.
- Evaluation of model performance on different “asset types” (e.g. securities, deposits, loans).
- Monitoring of production results: cross-check with the insurance corporations during a real data production round.

Thank you for your attention.

Vittoria La Serra vittoria.laserra@bancaditalia.it
Emiliano Svezia emiliano.svezia@bancaditalia.it