



BANCA D'ITALIA
EUROSISTEMA

Joint machine learning project on the list of assets template



*Joint workshop with experts from EIOPA, the ECB, NCAs, NCBs
and industry representatives on Insurance Reporting*

October, 20th 2021

Outline

- The joint project
- The data
- The problem to solve
- The aim of the project
- The proposed approach
- An application to Italian data
- Conclusions

The joint project

- **Period:** March - November 2021
- **Participants** from ECB, EIOPA, NCBs and NCAs
 - Only active contributors (small composition)
 - With observers (large composition)
- One meeting every two weeks (alternately small/large composition)
- **Coordinators**
 - by Banque de France (up to 1° October)
 - by Banca d'Italia (currently)

The data

- Data reported **quarterly** by ICs in the “list of assets” **Solvency II template SE.06.02 since 2016-Q1**.
- Each record is uniquely identified by a **corporation**, an **asset identification code** and an **observation date**.
- Qualitative and quantitative variables are observed on the records:
 - **Qualitative variables** include issuer/counterpart sector, issuer/counterpart area, type of asset (CIC), reporting sector of IC, original currency.
 - **Quantitative variable** of main interest is the assets’ market value (Solvency II amount).

The problem to solve

We expect the identification codes of the assets to remain unique and consistent over time.

However, due to the annual updates of the taxonomy or due to **reporting errors**, we can observe changes in the ID codes of the assets reported by ICs, from quarter to quarter.

Observation quarter	Insurance corporation	Asset ID code	Qualitative variables		Quantitative variables	
2019-Q1	A	X
2019-Q2	A	X
2019-Q3	A	X
2019-Q4	A	Y
2020-Q1	A	Y

The aim of the project

- The unexpected changes in the asset ID codes raise **issues when analyzing time series.**
- Our aim is:
 - to **develop a model**
 - that is capable of **automatically** identifying the cases of changes in the codes
 - searching for **couples of assets** that **don't share the same identification code but actually represent the same asset**
- **ICs will be involved** to check the potential anomalies identified by the model.

The proposed approach

A supervised machine learning (classification) approach

- Compare assets from two subsequent quarters, e.g. Q1 and Q2.
- Each couple of assets, one from Q1 and the other from Q2, can either be:
 - **Match** (the two assets share the same ID code)
 - **Non-match** (the two assets have different ID codes)
- **The goal:** to define a model that correctly classifies each couple as a match or non-match.
- Use machine learning techniques to train a supervised classification model.

The proposed approach

A supervised machine learning (classification) approach

Take two datasets with assets from two subsequent quarters Q1 and Q2.

Each asset can either be:

- Present in both quarters (e.g. X below)
- Present in Q1 and absent in Q2 (e.g. Y below)
- Absent in Q1 and present in Q2 (e.g. K below)

Dataset Q1			
Observation quarter	IC	Asset ID code	Other variables
Q1	A	X	...
Q1	A	Y	...
...

Dataset Q2			
Observation quarter	IC	Asset ID code	Other variables
Q2	A	X	...
Q2	A	K	...
...

The proposed approach

A supervised machine learning (classification) approach

Compare each couple of records, one from Q1 and the other from Q2, with respect to the observed qualitative variables and build a dummy matrix.

Asset ID code - Q1	Asset ID code - Q2	Insurance corporation	Qualitative variable 1	Qualitative variable 2	...	Qualitative variable N
X	Y	1	1	1	...	0
X	X	1	1	0	...	1
X	K	1	0	0	...	0
...

This field is blocked: **only compare the couples that share the same IC**

Qualitative variables in the two records are compared:
1 if the two records share the same value of qualitative variable, 0 otherwise

The proposed approach

A supervised machine learning (classification) approach

Run a supervised classification model on the dummy matrix, where the **target variable** is a binary one and can assume values in the set: {**the couple is a match**; **the couple is not a match**}.

	Asset ID code - Q1	Asset ID code - Q2
Non-match	X	Y
Match	X	X
Non-match	X	K

Insurance corporation	Qualitative variable 1	Qualitative variable 2	...	Qualitative variable N
1	1	1	...	0
1	1	0	...	1
1	0	0	...	0
...

Used as covariates in the model

An application to Italian data

Training and testing the model:

- Dataset Q1: a sample of assets from 2020-Q4
- Dataset Q2: a sample of assets from 2021-Q1
- In each quarter, around 100 insurance corporations and around 70,000 reported assets.
- Build the dummy matrix containing all the couples of assets from Q1 and Q2
- The dummy matrix is sampled to contain 95% non-matches and 5% matches
(out of the assets in the two quarters, we observed on average 5% of new assets)
- The best trained supervised classification model is *bagging* – a random trees based ensemble model

An application to Italian data

Results on the test set:

We evaluate results with respect to three metrics.

Metric	Description	Value
Accuracy	Percentage of correct predictions .	98.2%
Alpha error	Percentage of actual matching cases that were predicted as non-matches .	1.2%
False positive rate	Percentage of predicted matching cases that were actually non matches .	26.4%

Conclusions

Advantages of the proposed approach

- Machine learning models can increase effectiveness and efficiency of data quality checks
- Good preliminary results from the first applications

Limits of the proposed approach and possible next steps

- Not all the past information on the assets is used
- Only qualitative observed variables are compared
- A large false positive rate occurred in the results

Thank you

Vittoria La Serra

Vittoria.Laserra@bancaditalia.it

Emiliano Svezia

Emiliano.Svezia@ecb.europa.eu