

Stacking machine learning models for anomaly detection: comparing AnaCredit to other banking datasets

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

*PASQUALE MADDALONI, DAVIDE NICOLA CONTINANZA, **ANDREA DEL MONACO**, DANIELE FIGOLI,
MARCO DI LUCIDO, FILIPPO QUARTA, GIUSEPPE TURTURIELLO*

Stacking machine learning models for anomaly detection: comparing AnaCredit to other banking datasets

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

***The views and ideas expressed during the presentation are those of the speaker
and do not necessarily reflect the views and the ideas of the Bank of Italy***

PASQUALE MADDALONI, DAVIDE NICOLA CONTINANZA, *ANDREA DEL MONACO*, DANIELE FIGOLI,
MARCO DI LUCIDO, FILIPPO QUARTA, GIUSEPPE TURTURIELLO

Improving and maintaining the quality of collected data is a crucial task for official statistics

...which is also complex when data are *big* and granular



...making data quality management very challenging!

Modern methods such as *machine learning* and *big data analytics* may come to help.

Such methods have been already adopted by many central banks to enhance the data quality management.

Central banks can collect data by issuing regulations that define

- the information of interest
- the reporting population

 data about credit disbursement to the economy

There are three surveys that collect data on credit disbursement:

- the Balance Sheet Items (**BSI**)
- the Financial Reporting (**FinRep**)
- **AnaCredit**

The BSI is meant for monetary policy and it collects aggregated **information on assets and liabilities of the balance sheets** of the MFIs resident in the territory of the euro area Member States.

The key features of BSI data are the characteristics of the underlying contracts:

- type of instrument
- duration
- currency
- information on the borrowers (i.e., economic sector and residence)

The FinRep is a survey that collects **accounting information on assets, liabilities, equity and statement** of profit and loss for supervisory purposes.

FinRep data are broken down by

- accounting portfolio
- credit quality status
- type of instrument
- relevant characteristics of the counterparty (i.e., economic activity and residence)

AnaCredit stands for “**ana**lytical **credit** dataset”.

It is a dataset that collects detailed information on individual bank loans in the euro area.

Its approach is **loan-by-loan**.

Why AnaCredit?

The financial crisis of 2007-08 and the European debt crisis of 2009-10 showed that aggregate statistics are not sufficiently adequate to understand the underlying economic and financial developments.

The ECB has to be aware of, understand and monitor these developments.

Hence, it issued the Regulation (EU) 2016/867 on the collection of granular credit and credit risk data.

AnaCredit has been designed to deliver the necessary additional information for monetary policy and financial stability tasks.

It is important to improve and maintain the quality of AnaCredit data.

As AnaCredit collects granular information on phenomena that are covered by either BSI or FinRep in an aggregated manner, cross-checks between such surveys on a reporting agent-reference date basis may be performed to enhance data quality management.

The ECB and the ESCB have already deployed deterministic cross-checks BSI/AnaCredit and FinRep/AnaCredit: outliers are signaled when certain indicators exceed a pre-selected threshold. However, such a threshold is set to be the same for all reporting agents at every reference date.

We propose a non-deterministic approach which allows one to tailor such thresholds over

- the reporting agent
- the characteristics of both the instrument and the counterparty

a new approach: assumptions

Such a cross-checking approach is based on two assumptions.

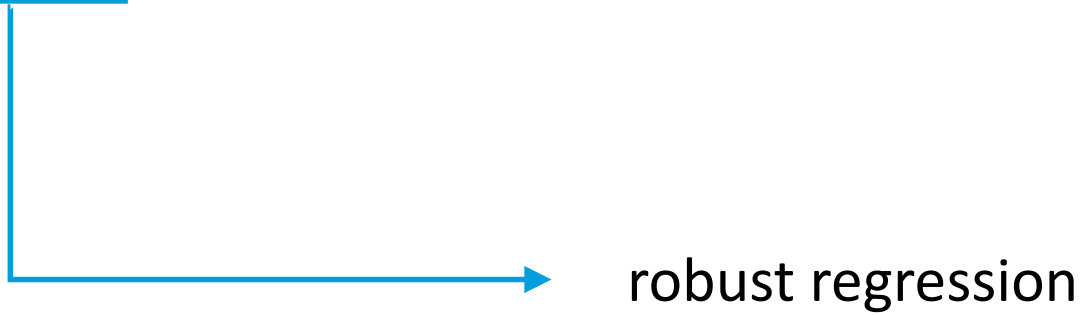
1. BSI, FinRep and AnaCredit contain similar information on loans
→ *patterns related to the same phenomenon are the same*
2. BSI and FinRep have been running for a longer period of time
→ *mismatches between the compared datasets should be attributed to anomalies in AnaCredit alone*

a new approach: methodology

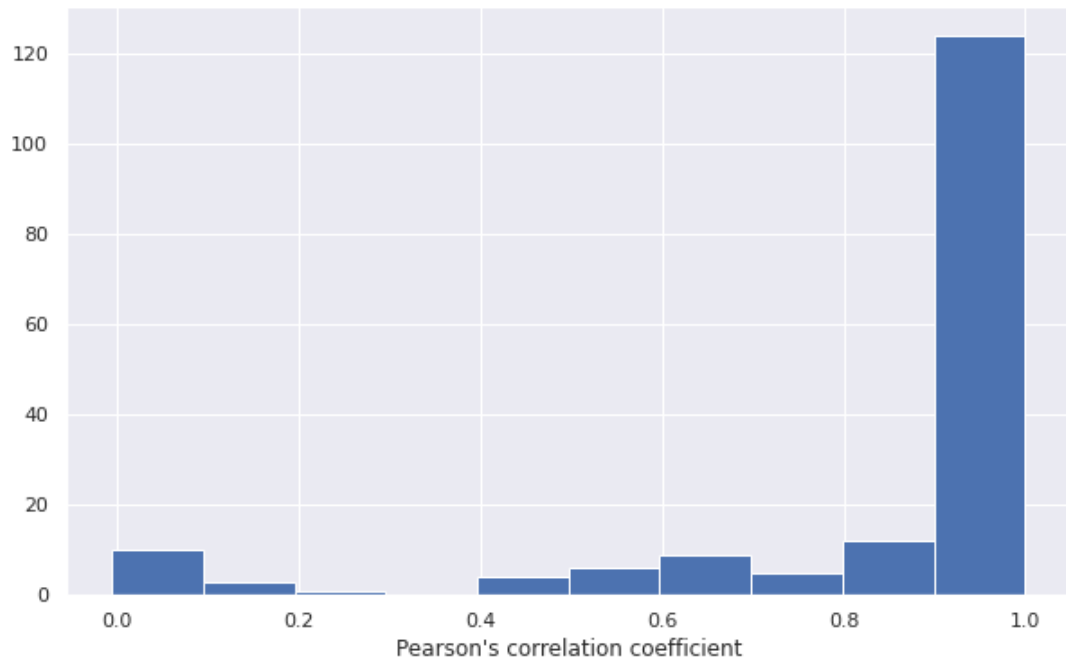
Supervised and unsupervised methods are combined via a stacking algorithm in a semi-supervised fashion.

a new approach: methodology

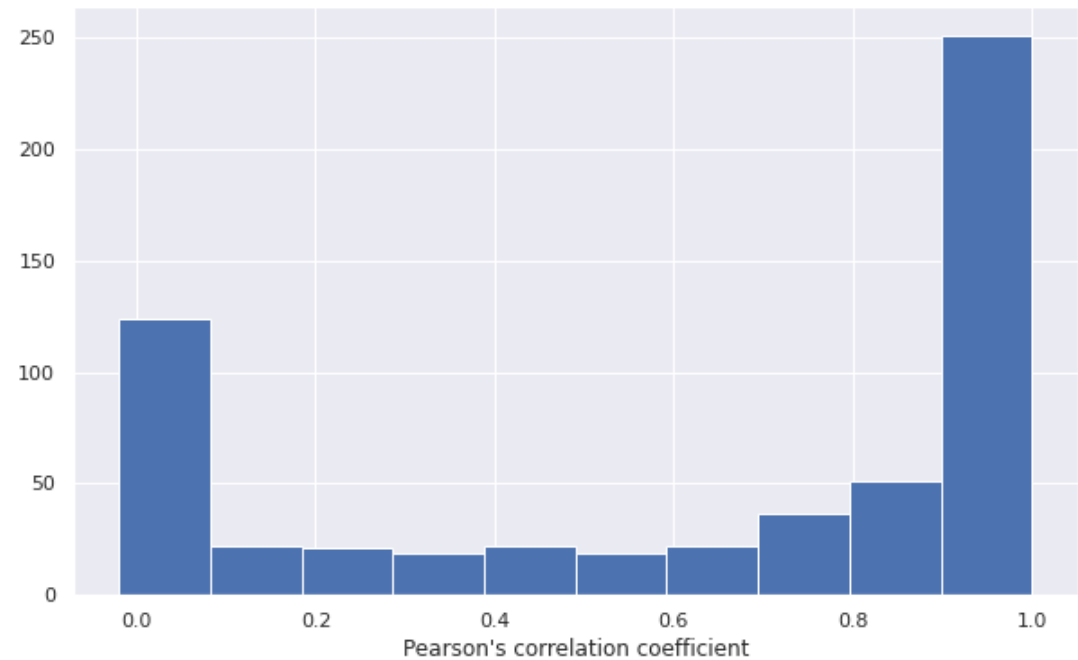
Supervised and unsupervised methods are combined via a stacking algorithm in a semi-supervised fashion.



BSI vs AnaCredit



FinRep vs AnaCredit



Supervised and unsupervised methods are combined via a stacking algorithm in a semi-supervised fashion.



```
graph TD; A[Supervised] --> B[robust regression]; B --> C["log(A_{i,j,t}) = beta_0 + beta_1 log(F_{i,j,t}) + beta_2 log(F_{i,j,t-1}/A_{i,j,t-1}) + epsilon_{i,j,t}"]
```

robust regression

$$\log(A_{i,j,t}) = \beta_0 + \beta_1 \log(F_{i,j,t}) + \beta_2 \log(F_{i,j,t-1}/A_{i,j,t-1}) + \epsilon_{i,j,t}$$

Supervised and unsupervised methods are combined via a stacking algorithm in a semi-supervised fashion.

robust regression

$$\log(A_{i,j,t}) = \beta_0 + \beta_1 \log(F_{i,j,t}) + \beta_2 \log(F_{i,j,t-1}/A_{i,j,t-1}) + \epsilon_{i,j,t}$$

AnaCredit amount


benchmark amount

explanatory variable
capturing structural differences

Supervised and unsupervised methods are combined via a stacking algorithm in a semi-supervised fashion.



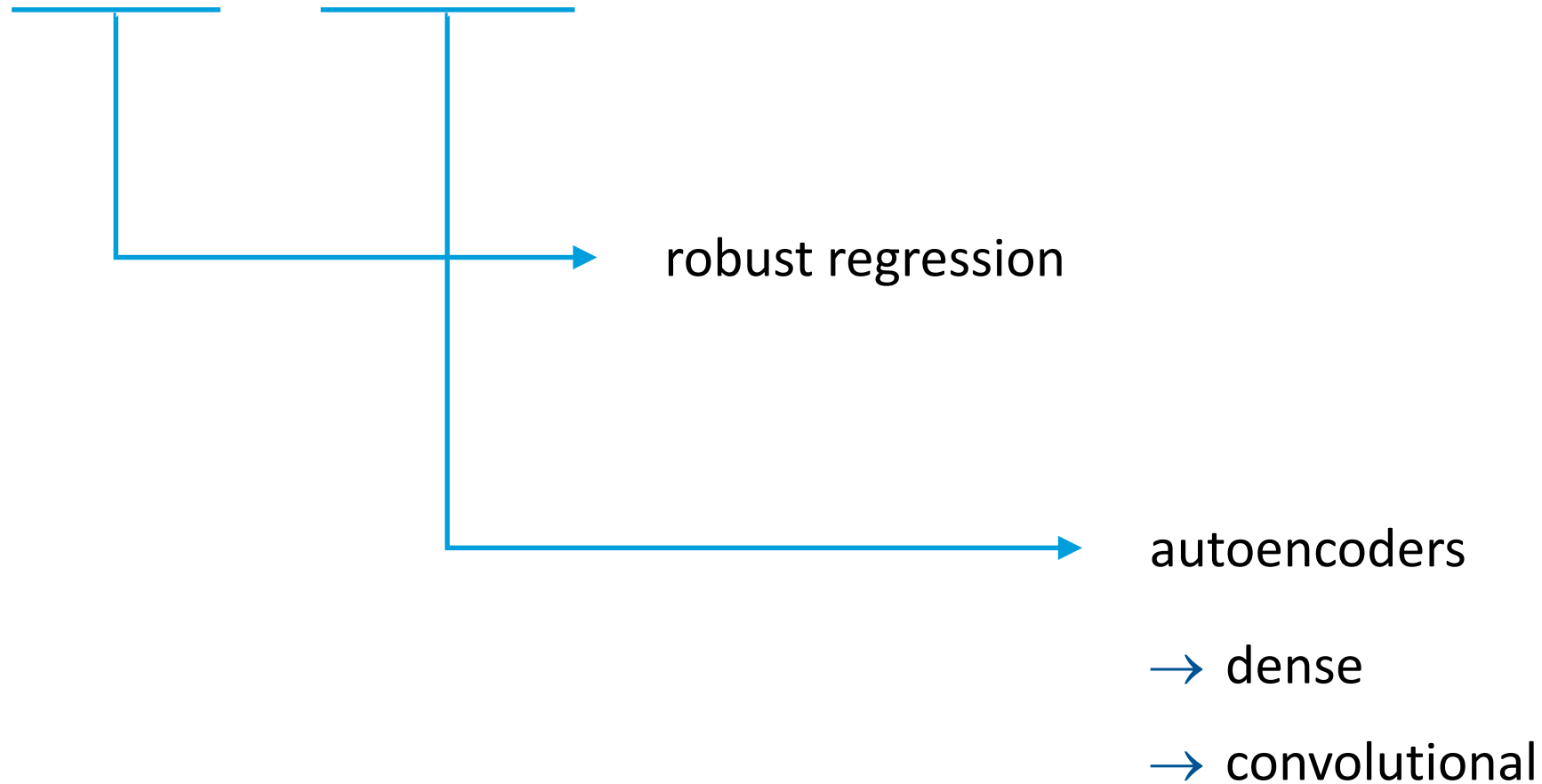
robust regression


$$\log(A_{i,j,t}) = \beta_0 + \beta_1 \log(F_{i,j,t}) + \beta_2 \log(F_{i,j,t-1}/A_{i,j,t-1}) + \epsilon_{i,j,t}$$

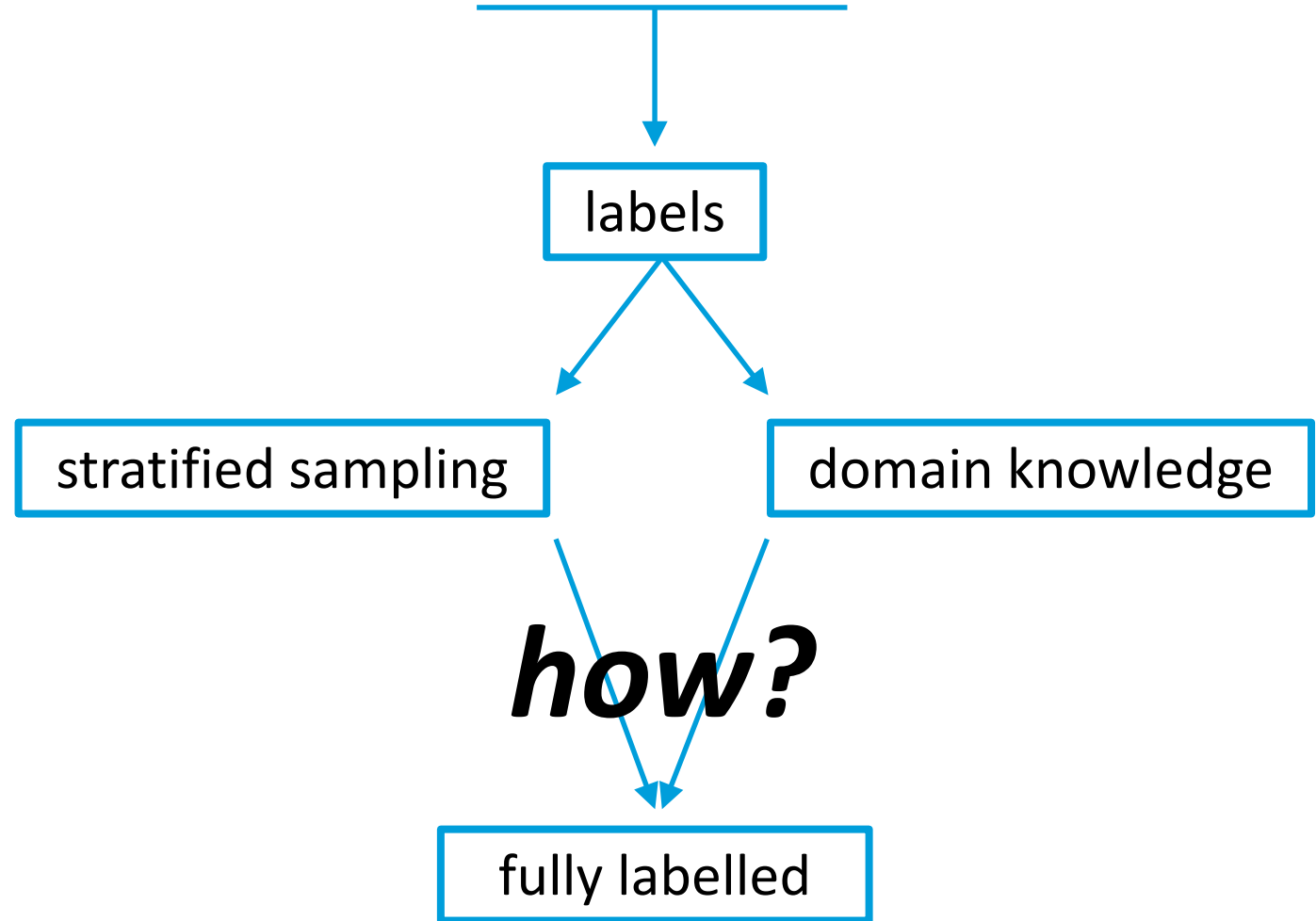
To solve the problem of exact fit:

- 1) *jittering*
- 2) standard robust regression
- 3) *thinning*

Supervised and **unsupervised** methods are combined via a stacking algorithm in a semi-supervised fashion.



Supervised and unsupervised methods are combined via a stacking algorithm in a semi-supervised fashion.

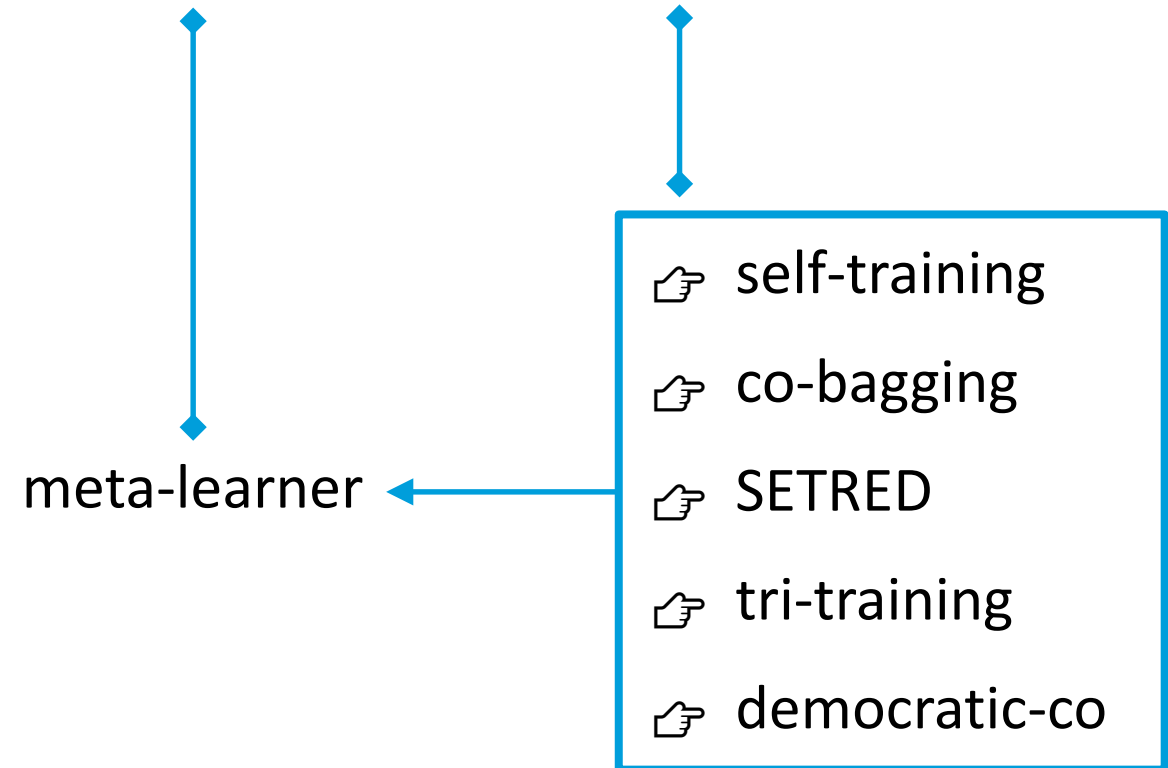


Supervised and unsupervised methods are combined via a stacking algorithm in a semi-supervised fashion.



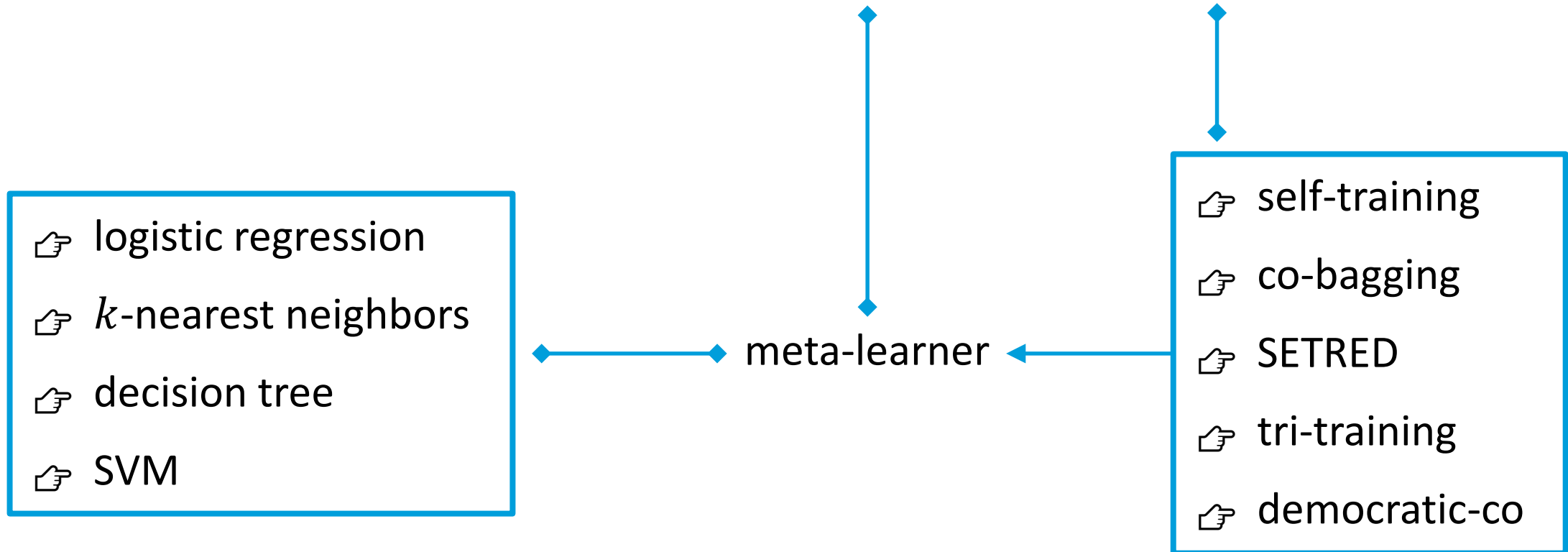
- 👉 self-training
- 👉 co-bagging
- 👉 SETRED
- 👉 tri-training
- 👉 democratic-co

Supervised and unsupervised methods are combined via a stacking algorithm in a semi-supervised fashion.

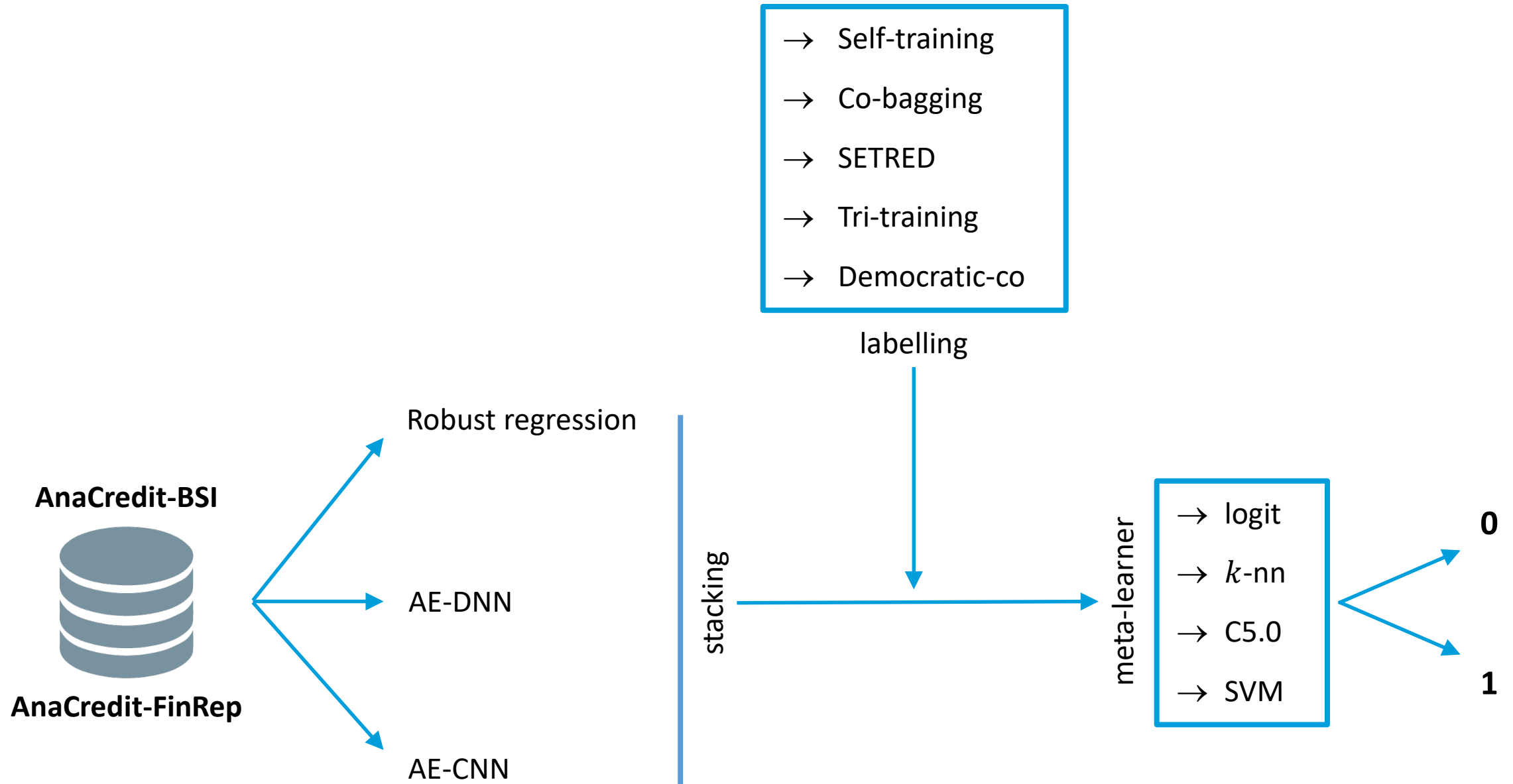


a new approach: methodology

Supervised and unsupervised methods are combined via a stacking algorithm in a semi-supervised fashion.



a new approach: the pipeline



a new approach: the results

- Self-training
- Co-bagging
- SETRED

AnaCredit-BSI

	Robust	AE-DNN	AE-CNN	Self-training	SETRED	Tri-training	Co-bagging	Democratic-co
<i>F1 score</i>	0.06109	0.04538	0.02632	0.995	0.991	0.991	0.992	0.992

Robust regression

AnaCredit-BSI

AnaCredit-FinRep

	Robust	AE-DNN	AE-CNN	Self-training	SETRED	Tri-training	Co-bagging	Democratic-co
<i>F1 score</i>	0.05296	0.00564	0.00564	0.598	0.634	0.634	0.650	0.500

inner

- logit
- k-nn

0

AnaCredit-FinRep

AE-CNN

1. The stacking technique outperforms the three selected base models
2. The proposed framework is flexible and general

Stacking machine learning models for anomaly detection: comparing AnaCredit to other banking datasets

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

thank you for your attention!

PASQUALE MADDALONI, DAVIDE NICOLA CONTINANZA, *ANDREA DEL MONACO*, DANIELE FIGOLI,
MARCO DI LUCIDO, FILIPPO QUARTA, GIUSEPPE TURTURIELLO