UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

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

Pasquale Maddaloni, Davide Nicola Continanza, Andrea del Monaco*, Daniele Figoli, Marco di Lucido, Filippo Quarta, Giuseppe Turturiello (Banca d'Italia, Italy)

*corresponding author: andrea.delmonaco@bancaditalia.it

I. Introduction[1]

1. Big-data analytics is increasingly being adopted within the community of central banks for several purposes (Cagala, 2017; Chakraborty and Joseph, 2017). An important area regards the application of machine learning techniques in order to improve the quality of data collected on the basis of regulatory reporting. Over the last few years, such surveys have become more granular and complex, in order to allow a better understanding of economic developments and, more in general, to improve the assessment of the actual and potential impact of policies on the economy[2]. As regards banking data, a key role is played by credit disbursement to the economy that, in Italy, represents more than two thirds of banks' total assets.
The main sources of credit data currently used at the Bank of Italy are the Eurosystem's collection of Balance Sheet Items (BSI), the EU harmonized Financial Reporting (FinRep), the Italian Central Credit Register data (CCR) and, for a couple of years now, the Eurosystem's granular collection AnaCredit.

3. This work investigates the possibility of building statistically founded cross-checking between the highly granular AnaCredit survey and the aggregated BSI and FinRep statistics by exploiting the similarities shared by the three surveys with respect to the phenomena that are covered. Originally, the three surveys were designed for different purposes; as such, the data collection framework is defined by different reporting rules and requirements on the types of loans that are collected, the reporting population, the data model and the preprocessing steps. Moreover, BSI and FinRep are very well established and mature surveys, whereas AnaCredit is a quite recent collection, so it might not have achieved the same high quality standards of the other two yet. This motivated us in exploiting the information available in BSI and FinRep to improve the quality of AnaCredit data through outlier detection techniques.

4. From a methodological point of view, the main idea is to resort to machine learning techniques (Bishop, 2011; Hastie *et al.*, 2001 and 2013) in order to carry out systematic cross-checks between the above mentioned datasets in order to identify potential outliers[3] and to overcome some of the limits recognized in the statistical literature on classical outlier detection, which are related both to the possibility that 'normal behaviour' might not be static but, rather, evolve over time and also to the lack of labelled data for training models (Chandola *et al.*, 2009). Within this research field (Cusano *et al.*, 2021; Zambuto *et al.*, 2020; Farnè *et al.*, 2018; Goldstein *et al.*, 2016), in this paper we discuss a general approach that can leverage ensemble techniques to perform pairwise comparisons between datasets containing information on similar phenomena. In particular, we show that such a methodology is able to detect anomalies with a higher level of precision than the single methods themselves used as baselines. Since anomalous observations are rare, the main metric considered to evaluate the performance of our developed models is the F1-score. In conclusion, we will show that the implementation of such a methodology can improve the quality of AnaCredit data, for the pairwise comparison against BSI and FinRep datasets can

[1] A longer version of this paper has been published in Occasional Papers, No. 689, Banca d'Italia, 2022

[2] For a recent discussion see Cœuré, 2017.

[3] As in Aggarwal, 2017, records will be considered anomalous if they are significantly different from the other points of the dataset

yield to a more accurate list of potential outliers to be submitted to the cross-checking of reporting banks. It is worth remarking how the approach developed in this paper can be applied, more generally, to all those cases in which there is the need to perform quality checks on information that are contained in both aggregated and granular datasets.

5. The paper is organized as follows. Section 2 describes the three datasets under consideration and the deterministic pre-processing treatment carried out in order to make it possible to compare the aggregated series available for each of them. Section 3 explores the different strategies considered for detecting outliers and illustrates the developed ensemble machine learning techniques within a semi-supervised setting. Section 4 presents the results of the proposed approach. Section 5 summarizes the main conclusions, outlining the advantages of the proposed method and the possible directions for future research.

6. The authors are grateful to Gianluca Cubadda and Alessio Farcomeni (University of Rome "Tor Vergata") and Francesca Monacelli and Roberto Sabbatini (Bank of Italy) for their useful comments and fruitful discussions on a preliminary draft of the paper.

7. The views and ideas expressed in this paper are those of the authors and do not necessarily reflect the views and ideas of the Bank of Italy.

II. Data

8. Within the European context, the Bank of Italy collects aggregated credit information by means of two "surveys"[4]: the monthly Balance Sheet Items (BSI) and the quarterly Financial Reporting (FinRep).

9. The BSI survey[5] is meant for monetary policy and it collects aggregated information on assets and liabilities of the balance sheets of eligible monetary financial institutions resident in the Euro area. The key features of BSI data are the characteristics of the underlying contracts (such as type of instrument, duration and currency) and specific information on the borrowers (for instance, economic sector and residence).

10. FinRep[6] is a harmonized 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 and by relevant characteristics of the counterparty such as institutional sector and economic activity.

11. Finally, the AnaCredit survey has been set up in accordance to the Regulation (EU) 2016/867 on the collection of granular credit and credit risk data (ECB/2016/13)[7], which the ECB issued following the global financial crisis of 2007-08 and the European debt crisis of 2009-10. The AnaCredit information focuses on the single credit instrument that is included in a contract stipulated between a credit institutions and a borrower (Di Noia and Moretti, 2020). The novelty of AnaCredit data with respect to both BSI and FinRep data is its loan-by-loan approach, which provides a detailed description[8] of every single loan granted to a legal entity.

III. Anomaly detection

12. As AnaCredit collects granular information on the same phenomena as covered by either BSI or FinRep in an aggregated manner, the quality of data can be enhanced by performing cross-checks between the three above mentioned surveys on a reporting agent-reference date basis.

13. The ECB and the National Central Banks (NCBs) have already developed deterministic cross-checks between BSI and AnaCredit: outliers are signaled when certain indicators exceed a pre-selected threshold. On

[4] From now on, by 'survey' we mean a collection of homogenous data under a reporting framework

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02013R1071-20141221

[6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014R0680

[7] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0867

[8] It counts over ninety features for every single reported financial instrument

the one hand, such a threshold is specific for each check; on the other, it is the same for each reporting agent and for every reference date.

14. In this paper we propose a non-deterministic approach which allows us to tailor such thresholds over the reporting agents and the characteristics of both the instrument and the counterparty at hand.
Such a cross-checking approach is based on the following assumptions.
Firstly, BSI, FinRep and AnaCredit contain similar information on loans; therefore, the patterns which are found in data referred to the same phenomenon are similar.
Secondly, since both BSI and FinRep have been running for a longer period of time with respect to AnaCredit, their quality is very high. Hence, potential outliers identified on the basis of mismatches in the compared datasets (namely, BSI vs. AnaCredit and FinRep vs. AnaCredit) can be attributed to anomalies in AnaCredit data alone.

IV. Methods

15. From the methodological point of view, we combine both supervised and unsupervised methods via a stacking algorithm[9] in a semi-supervised fashion. In this paper, the chosen supervised model is a robust regression, whereas the unsupervised models include two different implementations of autoencoders.

16. For sake of clarity, in what follows AnaCredit will be referred to as the “testing dataset” while any survey between the BSI and the FinRep will be referred to as the “benchmark dataset”. Since AnaCredit is compared against the BSI and the FinRep separately, we do not lose generality.

A. Robust regression

17. As the empirical evidences have shown the high correlation between aggregates from the testing dataset and their benchmark counterparties, a linear regression framework might apply.

18. In our model, the benchmark data is set as the independent variable, whereas the aggregated testing data is treated as the dependent variable. Furthermore, explanatory variables are introduced in order to capture structural differences between the two datasets. The resulting model is as follows:

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

where *i* denotes a bank, *j* a sub-portfolio of loans, *t* a reference date, $A_{i,j,t}$ the amount from the testing dataset and $F_{i,j,t}$ the amount from the benchmark dataset.
The above model takes into considerations only the differences at time *t-1* so that equation (1) can be read as an error correction model for the cross comparison of the two aggregates at time *t*.
It is worth mentioning that regression model in (1) does not include any seasonal variable, for AnaCredit is too young to have a long enough time series.
Notice also that the presence of anomalous points in the dependent variable yields to bad fitting, which forces us to appeal to statistical robust estimation (Hampel 1985; Hampel *et al.*, 1986; Farcomeni and Greco, 2015; Gschwandtner and Filzmoser, 2012; Maechler *et al.*, 2021). In particular, we consider the SMDM estimation proposed by Koller e Stahel (2011) that has both high asymptotic efficiency and high breakdown point (BP; see Hampel *et al.*, 1985).

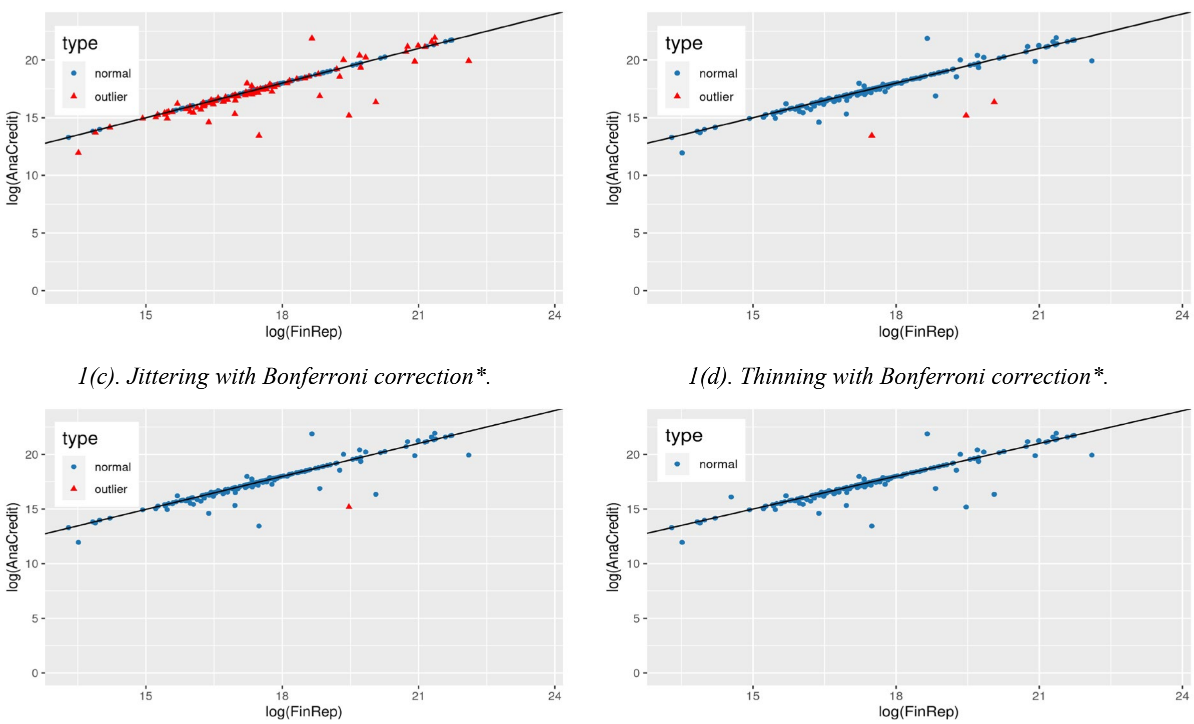
19. However, the problem of “exact fit” arises: points close to the regression line are classified as outliers by the SMDM algorithm. To solve this problem, we set a three stages procedure (see also Figure 1):

(a) we add Gaussian noise to the testing data (i.e., *jittering*) and run the regression; if the null hypotheses on $\widehat{\beta_0}$, $\widehat{\beta_1}$, and $\widehat{\beta_2}$ are not rejected and the R^2 is above the 80%, observations are signaled as outliers in accordance to the χ^2 test on residuals with the Bonferroni correction, otherwise

(b) we run the standard robust regression and check whether the null hypotheses on $\widehat{\beta_0}$, $\widehat{\beta_1}$, and $\widehat{\beta_2}$ are not rejected and the R^2 is above the 80%; if this is the case, observations are signaled as outliers in accordance to the χ^2 test on residuals with the Bonferroni correction, otherwise

[9] Roughly, the stacking technique consists in training a “meta-learner” on the outputs of other models to get a final prediction

Figure 1.



1(a). No correction.

1(b). Bonferroni correction.*

1(c). Jittering with Bonferroni correction.*

1(d). Thinning with Bonferroni correction.*

* $\alpha = 0.05$

(c) we perform a *thinning* procedure by removing points as described in Cerioli and Perrotta, 2013; if the null hypotheses on $\widehat{\beta_0}$, $\widehat{\beta_1}$, and $\widehat{\beta_2}$ are not rejected and the R^2 is above the 80%, observations are signaled as outliers in accordance to the χ^2 test on residuals with the Bonferroni correction.

B. Autoencoders

20. An autoencoder is a neural network model that is able to perform hierarchical and nonlinear dimensionality reduction of data. The goal of an autoencoder is to learn structural patterns of data by setting the target as its own input. Generally, its architecture is symmetric: both the input and the output layers have the same number of nodes, that decrease and increase while moving forward through the hidden layer nodes.

21. Formally, assuming that data lies in n-dimensional space R^n, autoencoders aim at reconstructing the data through a bottleneck: they push the input down to a k-dimensional space R^k with $k \ll n$ (the so-called "latent space") via a function $E: R^n \to R^k$ defined as the "encoder", then they pull the result back to the input space R^n via a function $E: R^k \to R^n$, called "decoder". Hence, the autoencoding process be expressed as

$$\hat{x} = D\big(E(x)\big) =: D(z), \tag{2}$$

where z is the latent vector and $\hat{x}$ is the reconstructed input x. It is worth noticing that autoencoders are unsupervised models, for they do not need labels since the target variable is the input itself.

22. In this paper, we propose two different autoeconders: the convolutional autoencoder (AE-CNN) and the dense autoencoder (AE-DNN). In the AE-CNN, the encoder's architecture E consists of three hidden layers with convolutional filters of respectively 16, 32, 16 units, kernel sizes of 3x3 and a max-pooling window of 2x2. The bottleneck consists of a convolutional layer with 4 convolutional filters of 23x23 size. The architecture of the decoder D mirrors the encoder's one, with up-sampling substituting the max-pooling. In the AE-DNN, both the encoder E and decoder D networks consist of 2 fully-connected layers with respectively 150 and 28 hidden units, with Leaky ReLU as activation function. The first layer also contains a dropout (Srivastava *et al.*, 2014) of 5%. The bottleneck layer is chosen to be a fully connected layer with 20 hidden units, which results in a 20-dimensional latent space. To measure the reconstruction accuracy, we make use of the Structural Similarity Index Metric (SSIM), as in Wang *et al.*, 2004.

23. Thanks to their ability to learn a reduced representation within a latent space, autoencoders can be fruitfully employed to detect outliers in data (Russo *et al.*, 2019). The idea is that outliers are not well represented in the latent space, so the error of outliers' reconstruction will be larger and, therefore, better identifiable.
In this paper, we propose to train the autoencoders on the benchmark dataset and to apply them to the testing dataset.

C. Semi supervised labelling and the stacking technique

24. The three above mentioned models produce an anomaly score for each point in the testing dataset. The robust regression scores show low correlation with the autoencoders' ones. Moreover, scores generated by AE-CNN and AE-DNN are not remarkably correlated. Hence, we can move forward by stacking such scores to produce the final prediction (Wolpert, 1992; Dzeroski and Zenko, 2004).

25. However, stacking models require the presence of labels, *i.e.* a response variable that has to be matched. To get a response variable, some cases have been sampled in a stratified manner by appealing to the Neyman optimal criterion (Neyman, 1934) and subsequently verified with reporting agents. Some others cases have been pre-labelled on the basis of the domain knowledge of the analysts. Such information is bound to learners' scores, obtaining a final partially labelled dataset on which we can train meta-learners in a semi-supervised manner (Chapelle *et al.*, 2006; González *et al.*, 2019).

26. To get a fully labelled dataset, we take two different approaches. The first approach consists in completing the labelling procedure via a Monte Carlo simulation. Pseudo labels (i.e., simulated responses) to the not sampled and not pre-labelled observations are assigned randomly by replicating the sample distribution of the responses received by reporting agents within each stratum. The second approach consists in applying algorithms that can learn by themselves how to label unlabelled observations. The chosen algorithms to perform such a task are: Self-training (Yarowsky, 1995), SETRED (Li and Zhou, 2005), Tri-training (Zhou and Li, 2005), Co-Bagging (Blum and Mitchell, 1998) and Democratic-Co (Zhou and Goldman, 2004). Pairwise comparisons of such models are carried out by the nonparametric McNemar's Test for binary classification (see Table 1 and Table 2).

Table 1. AnaCredit-BSI cross-checks: predictions comparison of the semi-supervised learning algorithm in terms of p-value.*

	Self-training	SETRED	Tri-training	Co-Bagging	Democratic-Co
Self-training		0.009	0.138	0.003	0.013
SETRED			0.269	0.000	0.251
Tri-training				0.002	0.025
Co-Bagging					0.000
Democratic-Co					

* a p-value lower than 0.05 indicates a significant disagreement between the two models predictions

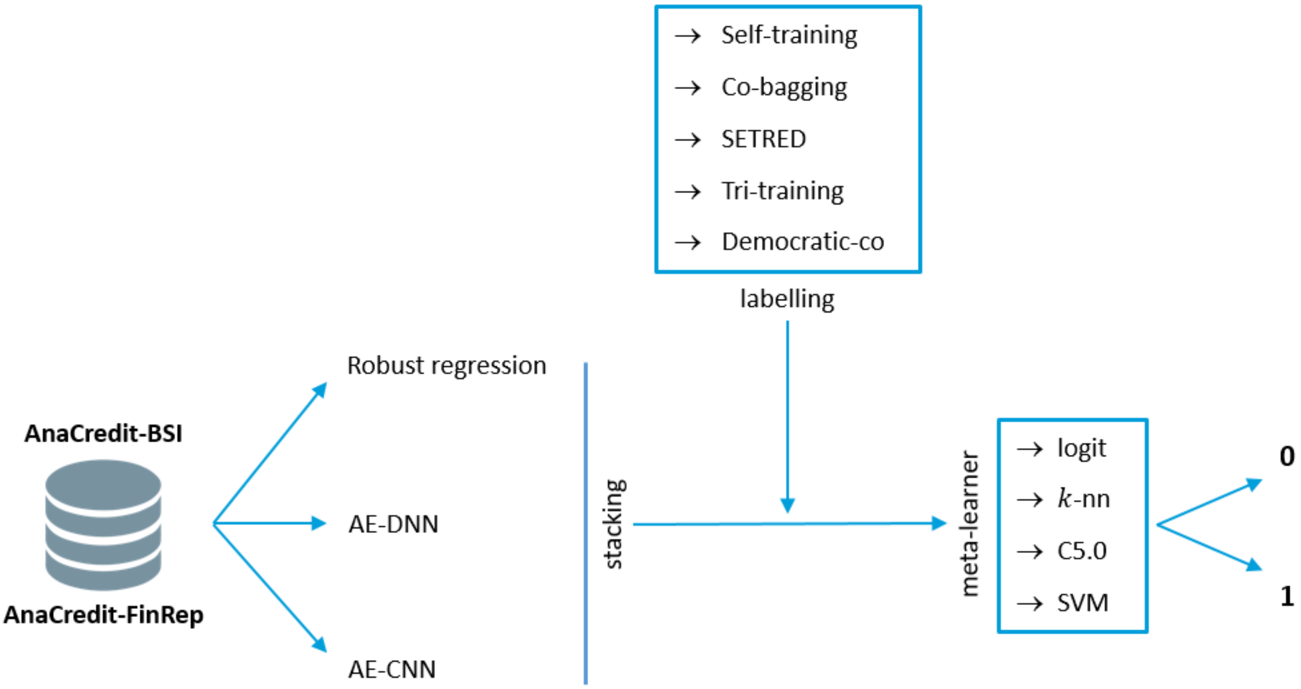
Table 2. AnaCredit-FinRep cross-checks: predictions comparison of the semi-supervised learning algorithm in terms of p-value.*

	Self-training	SETRED	Tri-training	Co-Bagging	Democratic-Co
Self-training		0.473	0.787	0.833	0.080
SETRED			0.488	0.573	0.184
Tri-training				0.500	0.089
Co-Bagging					0.087
Democratic-Co					

* a p-value lower than 0.05 indicates a significant disagreement between the two models predictions

27. Finally, a meta-learner is trained to the fully labelled dataset of scores. The models considered as meta-learners are a logistic regression, a *k*-nearest neighbor, a decision tree (namely, the C5.0 algorithm) and a support vector machine (SVM). The final pipeline is graphically represented in the Figure 2 below.

Figure 2. The pipeline with the semi-supervised labelling approach.



V. Results and conclusions

28. The stacking technique outperforms the three selected base models (Robust Regression, AE-CNN, AE-DNN) as it allows us to combine the complementary insights that stem from their outputs[10]. The results are shown in Table 3 and Table 4 below.

Table 3. AnaCredit-BSI cross-checks: performances of the base learning models against the SVM meta-learner on semi-supervised labelled scores.

	RobReg	AE-DNN	AE-CNN	Self-training	SETRED	Tri-training	Co-Bagging	Democratic-Co
F1 score	0.061	0.045	0.026	0.995	0.991	0.991	0.992	0.992

Table 4. AnaCredit-FinRep cross-checks: performances of the base learning models against the C5.0 meta-learner on semi-supervised labelled scores.

	RobReg	AE-DNN	AE-CNN	Self-training	SETRED	Tri-training	Co-Bagging	Democratic-Co
F1 score	0.053	0.006	0.006	0.598	0.634	0.634	0.650	0.500

29. In particular, we notice that the SVM meta-learner trained on the scores dataset that has been fully labelled by the self-training algorithm achieves the higher F1-score in AnaCredit-BSI cross-checks, whereas the Co-Bagging model together with meta-learner C5.0 yield the best performance in the AnaCredit-FinRep comparison.

30. As for the approach to get the fully labelled dataset of scores, we noticed that the Monte Carlo simulation did not yield very remarkable performances as those achieved by the semi-supervised algorithms.

[10] Lessmann et al., 2015

31. It is worth noticing that many refinements of the current work are possible. For instance, one might appeal to a panel regression in place of the robust regression or to variantional autoencoders rather than standard autoencoders. Nevertheless, the framework we propose in this paper is flexible and general, and can be applied to carry out pairwise comparisons between any collection of datasets containing information on similar phenomena but with different levels of granularity.

VI. References

Aggarwal C. (2017). "Outlier Analysis", Springer.

Bishop, C.M. (2011). "Pattern Recognition and Machine Learning", Springer.

Blum A. and Mitchell T. (1998). "Combining labeled and unlabeled data with co-training", in Eleventh Annual Conference on Computational Learning Theory, COLT' 98, pages 92–100, New York, NY, USA.

Cagala, T. (2017). "Improving Data Quality and Closing Data Gaps with Machine Learning", IFC Bulletin, 46.

Cerioli A. and Perrotta D. (2013). "Robust clustering around regression lines with high-density regions", Springer, Advances in Data Analysis and Classification volume 8, pp. 5–26.

Chakraborty C. and Joseph A. (2017). "Machine Learning at Central Banks", Bank of England Staff Working Paper No. 674, https://doi.org/10.2139/ssrn.3031796

Chapelle O., Scholkopf B. and Zien A. (2006). "Semi-supervised learning", MIT Press

Chandola V., Banerjee A. and Kumar V. (2009). "Anomaly detection: a survey", ACM Computing Surveys, Vol. 41, No. 3, http://doi.acm.org/10.1145/1541880.1541882

Cœuré B. (2017). "Setting standards for granular data", Opening remarks by Benoît Cœuré, Member of the Executive Board of the ECB, at the Third OFR-ECB-Bank of England workshop on "Setting Global Standards for Granular Data: Sharing the Challenge", Frankfurt am Main, 28 March 2017, https://www.ecb.europa.eu/press/key/date/2017/html/sp170328.en.htm

Cusano F., Marinelli G. and Piermattei S. (2021*)*. "Learning from revisions: a tool for detecting potential errors in banks' balance sheet statistical reporting", Bank of Italy, Working Papers, No. 611.

Dzeroski, S. and Zenko, B. (2004). "Is combining classifiers with stacking better than selecting the best one?", Machine Learning, 255–273.

Di Noia M. and Moretti D. (2020). "Le informazioni statistiche della Banca d'Italia sul rischio di credito e la nuova rilevazione AnaCredit", Banca d'Italia, Occasional Papers, No. 544.

Farcomeni A. and Greco L. (2015). "Robust methods for data reduction", CRC Press.

Farnè M. and Vouldis A.T. (2018). "A methodology for automatised outlier detection in high-dimensional datasets: an application to euro area banks' supervisory data", ECB Working Paper N. 2171.

Goldstein M. and Uchida S. (2016). "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data", Computer Science, Medicine, PLoS ONE.

González M., Rosado O., Rodríguez J. D., Bergmeir C., Triguero I. and Benítez J. M. (2019). "ssc: An R Package for Semi-Supervised Classification", R package version 2.1-0.

Hampel, F. R. (1985). "The Breakdown Point of the Mean Combined With Some Rejection rules", Technometrics, 27, 95-107.

Hampel, F., Ronchetti E., Rousseeuw P. and Stahel W. (1986). "Robust Statistics: The Approach Based on Influence Functions", N.Y.: Wiley

Hastie T., Tibshirani R. and Friedman J. (2001). "The Elements of Statistical Learning", Springer.

Hastie T., James G., Tibshirani R. and Witten D. (2013). "An Introduction to Statistical Learning", Springer.

Koller, M. and Stahel W. A. (2011). "Sharpening wald-type inference in robust regression for small samples", Computational Statistics & Data Analysis 55(8), 2504–2515

Lessmann S., Baesens B., Seow H.V. and Thomas L.C. (2015). “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research”. *European Journal of Operational Research*, 247 (1), 124-136.

Li M. and Zhou Z. (2005). “Setred: Self-training with editing. In Advances in Knowledge Discovery and Data Mining”, volume 3518 of Lecture Notes in Computer Science, pages 611–621. Springer Berlin Heidelberg.

Maechler M., Rousseeuw P., Croux C., Todorov V., Ruckstuhl A., Salibian-Barrera M., Verbeke T, Koller M., Conceicao E.L. and Anna di Palma M. (2021). “robustbase: Basic Robust Statistics”, R package version 0.93-7, http://robustbase.r-forge.r-project.org/

Maronna*, R.A.*, Martin, D.R. and Yohai, V.J. (2006). “Robust Statistics: Theory and Methods”, Wiley, New York.

Neyman, *J. (1934).* “On the two different aspects of the representative methods. The method stratified sampling and the method of purposive selection”, Journal of Royal Statistical Society, 97, 558-606.

Russo S., Disch A., Blumensaat F. and Villez K. (2019). “Anomaly detection using deep autoencoders for in-situ wastewater systems monitoring data”. Proceedings of the 10th IWA Symposium on Systems Analysis and Integrated Assessment (Watermatex2019), Copenhagen, Denmark, September 1-4.

Srivastava N, Hinton G., Krizhevsky A., Sutskever I. and Salakhutdinov R. (2014). “Dropout: a simple way to prevent neural network from overfitting”, Journal of Machine Learning Research, 15.

Zambuto F., Buzzi M. R., Costanzo G., Di Lucido M., La Ganga B., Maddaloni P., Papale F. and Svezia E. (2020). “Quality checks on granular banking data: an experimental approach based on machine learning”, Banca d’Italia, Occasional Papers, No. 547.

Zambuto F., Arcuti S., Sabatini R. and Zambuto D. (2020). “Application of classification algorithms for the assessment of confirmation to quality remarks”, Banca d’Italia, Occasional Papers, No. 631.

Zhou and Goldman S. (2004). “Democratic co-learning”, in IEEE 16th International Conference on Tools with Artificial Intelligence (ICTAI), pages 594–602.

Zhou Z. and Li M. (2005). “Tri-training: exploiting unlabeled data using three classifiers”, IEEE Transactions on Knowledge and Data Engineering, 17(11):1529–1541.

Wang Z., Bovik A.C., Sheikh H.R. and Simoncelli E.P. (2004). “Image quality assessment: from error visibility to structural similarity”, IEEE transactions on image processing, 13(4):600–612.

Wolpert, D. (1992). “Stacked generalization”, Neural Networks, 5, 241-260, https://doi.org/10.1016/S0893-6080(05)80023-1.

Yarowsky D. (1995). “Unsupervised word sense disambiguation rivaling supervised methods”. In Proceedings of the 33rd annual meeting on Association for Computational Linguistics, pages 189–196, Association for Computational Linguistics.