11th Biennial IFC Conference on "Post-pandemic landscape for central bank statistics"

BIS Basel, 25-26 August 2022

A decision-making rule to detect insufficient data quality – an application of statistical learning techniques to the non-performing loans banking data[1]

Paolo Cimbali, Marco De Leonardis, Alessio Fiume, Barbara La Ganga,
Luciana Meoli and Marco Orlandi,
Bank of Italy

[1] This presentation was prepared for the conference. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the event.

A decision-making rule to detect insufficient data quality

an application of statistical learning techniques to the non-performing loans banking data

Barbara La Ganga, Paolo Cimbali, Marco De Leonardis, Alessio Fiume, Luciana Meoli and Marco Orlandi

Abstract

The data quality level generally follows an improving trend thanks to subsequent corrections submitted by reporting agents; however, a data quality worsening may occur when data production is affected by exogenous and unpredictable events, such as the pandemic or changes in the reporting requirements. Using banks Non-performing loans regulatory requirements, the paper proposes a decision-making rule to improve the detection of these cases by defining a synthetic data quality indicator based on past evidence from data quality management activity: outliers, their severity level and received confirmations of underlying data, the latter considered for estimating the expected confirmations through a logistic regression model.

Keywords: outliers, non-performing loans, data quality, supervised machine learning, logistic regression.

JEL classification: C18, C81, G21.

Contents

1. Introduction 2

2. From data collection to the release of the information 3

3. The NPL dataset and model selection 6

4. Main results for the selected model 8

4.1 Logistic regression: estimation of 'remark confirmed, Yes/No' 8

4.2 Application of the decision-making rule 9

Conclusions 11

References 12

1. Introduction[1]

The Bank of Italy collects a large array of statistical and supervisory data from banks and other financial institutions on a regular basis to support its analyses and policy decisions. The data are organized in datasets, also called reports, each obeying different reporting regulations. Assessing the Data Quality Level of reports (DQL) is key to enabling users to carry out thorough and robust analyses.

Reporting Agents (RAs) are required to ensure high-quality data. Errors need to be promptly corrected by the RAs with a new report submission, whereas the unusual, yet correct, data have to be confirmed by the RAs. A Data Manager is in charge of evaluating the received confirmations and establishing a dialogue with the RAs in order to collect further information on unusual data patterns. This process creates a data quality cycle where RAs are required to submit new reports as long as the DQL is not sufficient (Casa *et al.*, 2022).

Through subsequent data submissions, DQL is expected to improve over time; however, the data reliability may be impaired by outliers that are actual errors. A worsening in the DQL could occur especially when data production is affected by exogenous and unpredictable events. These events include RAs' IT malfunctions in the production of statistical reports, errors occurring when applying changes in the reporting requirements, or extraordinary events, such as the pandemic, that can affect the organisation of the RAs (Schnabel, 2020; Tissot and De Beer, 2020). Prompt detection of a substantial DQL worsening is key for reducing costs in the Data Quality Management (DQM) process and getting fit-for-use data.

In literature on data quality, this topic is typically approached by monitoring some stand-alone dimensions, such as timeliness, accuracy and consistency, all evaluated through metrics and indicators (Damia and Aguilar, 2006; Pipino *et al.*, 2002). In the international statistical context, some Institutions have developed specific strategies and frameworks (IMF Data Quality Assessment Framework, 2012; ESS QAF European Commission and Eurostat, 2019). A peculiar approach to DQM is envisaged by the application of the Benford's law, also known as the 'first digit law'. This approach allows identifying errors as elements that do not follow an empirical regularity in data describing naturally-occurring phenomena (Gonzalez-Garcia and Pastor, 2009).

In this paper, with reference to the accuracy, completeness and consistency of reports sent by RAs, we investigate the improvement of the DQL monitoring by defining a new decision-making rule which could support the Data Manager in assessing whether a revision by an RA of previously transmitted data improves or worsens the DQL. The rule is based on a synthetic data quality indicator computed through past evidence from the DQM activity: the outliers, their severity level and the received confirmations of underlying data. These three metrics are considered to evaluate the DQL; furthermore, since the confirmation status is unknown once a new data submission is received, the confirmations related to remarks just generated are

[1] The authors are grateful to Gianluca Cubadda and Alessio Farcomeni (University of Tor Vergata, Rome), Laura Mellone, Francesca Monacelli and Roberto Sabbatini (Bank of Italy) for their useful comments and fruitful discussions on a preliminary draft of the paper. The views expressed herein are those of the authors and do not necessarily reflect those of the Bank of Italy.

estimated through statistical learning techniques using the information on the past DQM activity.

The methodology, applied to the Non-Performing Loans (NPL) dataset collected by the Bank of Italy, shows an improvement in the DQL monitoring process.

The paper is organized as follows. Section 2 describes the logical path that leads from the DQM process to the definition of the decision-making rule. Section 3 describes the data used in the empirical part of the study and illustrates the main arguments that guided the choice of the final model by comparing different alternatives. Section 4 shows the empirical results also validating the consistency of the decision rule with the data corrections that take place in practice.

2. From data collection to the release of the information

In the Bank of Italy the process spanning from the collection to the dissemination of monetary and financial data is structured in steps. As soon as a dataset is submitted by a RA, it undergoes the application of a set of automatic data quality checks (DQCs) to detect anomalies. When an outlier is identified, a notification (called 'remark') is generated and transmitted to the RA for its assessment and possibly for action to be taken.

Depending on the type of DQC, the generated remarks can request either a confirmation or a correction of the original data. DQCs that detect only reporting errors are defined as 'non-confirmable' since the data related to remarks generated ('non-confirmable remarks') cannot be confirmed. Each DQC is characterised by a 'severity level' defined in advance according to a scale increasing from 0 to 10 based on an *a priori* evaluation by data quality experts of the impact of errors on the quality of the underlying data. According to the 'severity' of DQCs, the generated remarks are also classified as 'serious remarks' and 'non-serious remarks', respectively. In particular, the 'serious remarks' are characterised by a severity level equal to 9 or 10; the 'non-serious remarks' by a severity level below 9.

RAs are required to assess outliers and to correct or confirm them by providing a suitable explanation for the underlying anomalous pattern. Based on the severity of the outstanding outlier as well as of the analysis of the explanations provided for the confirmations, the Data Manager may conclude that the DQL is still not adequate and therefore continue the dialogue with the RA open until the DQL reaches an overall satisfactory level. At each iteration, the Data Manager faces a trade-off: on the one hand, there is the need to make the information promptly available to users which implies that the above data quality cycle must be kept short; on the other hand, there is the need for the data to be as free as possible from significant errors. This assessment represents the core issue of the present study.

In detail, a RA submits the k^{th} dataset related to a specific reference date t. After the first round of DQCs, the RA receives a set of remarks by the Data Manager. After checking the evidence in its internal systems, the RA corrects the erroneous data by submitting the $(k+1)^{th}$ report. In so doing, the revisions can either enhance or, in some extreme but still plausible cases, introduce a new error which worsens the DQL. Then, the Data Manager faces the situation illustrated in Graph 1.

Overall quality from the k^{th} to the $(k+1)^{th}$ submission

Graph 1

		$(k+1)^{th}$ submission	
		Not-released	Released
k^{th} submission	Not-released	**D**	**C**
	Released	**A**	**B**

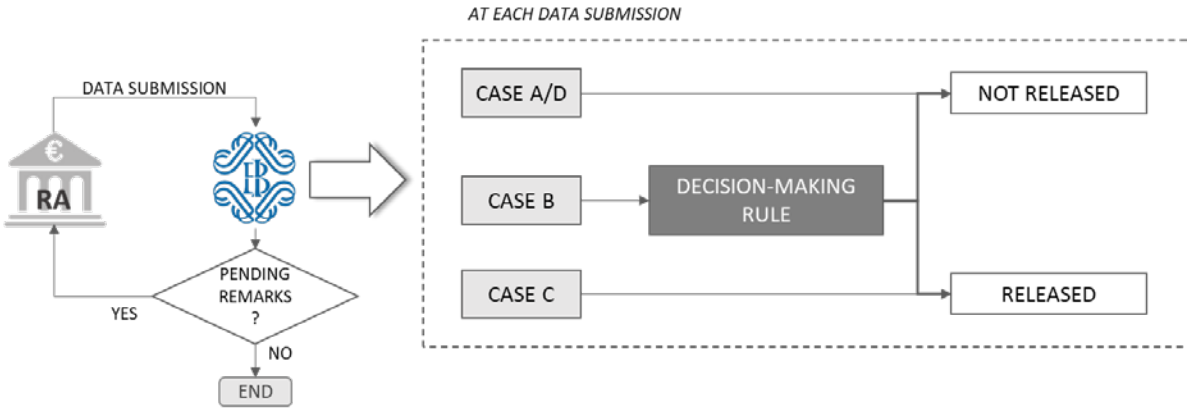
- CASE A: the k^{th} data submission is considered fit-for-use since the automatic DQCs have generated only non-serious remarks and information can be disseminated to the users. With the $(k+1)^{th}$ data submission of the same dataset, the new data worsen the DQL because of the presence of serious remarks, which prevent data to be made available to users.
- CASE B: both k^{th} and $(k+1)^{th}$ data submissions are considered fit-for-use because the automatic DQCs have generated only non-serious remarks. In this case, the latest data are always available to the users.
- CASE C: this event is the inverse situation of CASE A and it shows an improvement of DQL between the k^{th} and $(k+1)^{th}$ data submissions. Only the latest data are released to the users.
- CASE D: both k^{th} and $(k+1)^{th}$ data submissions are not released to the users because of the presence of serious remarks. The Data Manager and the RA carry on in their interaction in order to resolve the data issues.

In cases A, C and D, the Data Manager's decision is straightforward. Since in case B the $(k+1)^{th}$ data submission could contain some anomalies such that they reduce the DQL in comparison with the k^{th} submission, it is key that the Data Manager has an instrument to early detect the worsening of the DQL from one report to the next.

Graph 2 depicts the DQM process in presence of a decision rule based on the change of the DQL.

DQM - flow chart

Graph 2



In brief, the decision-making algorithm must mirror the DQL assessment made by the Data Manager when a subsequent data submission is transmitted by a specific RA and reference date. Then, we consider the DQL variation between two consecutive reporting submissions, the k^{th} and the $(k+1)^{th}$, of the same dataset ($\Delta DQL_{k+1}=DQL_{k+1}- DQL_k$) when both submissions do not bear serious remarks and are in principle fit-for-use. Hence, the rule needs to take as inputs the number of remarks generated at each round of DQCs, the severity level for each of them and the number of confirmed remarks.

Let us define the dummy variable *R* (remark) taking the value 1 if the DQC *c*, applied at the reference date *t* for the k^{th} data submission sent by the RA *p*, is violated and 0 otherwise

$$R_{t,p,c,k}=\begin{cases} 1, & \text{if c is violated} \\ 0, & \text{if c is satisfied} \end{cases}$$

and the dummy variable *Conf* taking the value 1 if the remark *R* is confirmed by the RA and 0 otherwise

$$Conf_{t,p,c,k}=\begin{cases} 1, & \text{if the } R_{t,p,c,k} \text{ is confirmed} \\ 0, & \text{otherwise} \end{cases}$$

In case of non-confirmable remarks, *Conf* is by construction equal to zero.

The synthetic data quality indicator for the k^{th} data submission is defined as follows:

$$I_{t,p,k} = \sum_{c=1}^{C} \tau_c \left(R_{t,p,c,k}\text{-}Conf_{t,p,c,k}\right)$$

where *C* denotes the number of remarks generated and τ_c represents the severity level of the DQC.

The higher the value of $I_{t,p,k}$, the lower the DQL_k. In particular, when the k^{th} data submission is not impaired by remarks not confirmed, the indicator is equal to zero.

Our decision rule (1) is such that the release of the $(k+1)^{th}$ data submission takes place when

$$I_{t,p,k+1} \leq I_{t,p,k} \quad (1)$$

The decision rule (1) assumes that the evidence of a confirmed remark is already known. However, the submission of an explanation from an RA is subsequent to the generation of the related remark, then this status, named $\widehat{Conf}$, has to be forecasted through the estimation of the probability *p(Conf)* that a remark is confirmed by a given RA, on a specific reference date:

$$\widehat{Conf}_{t,p,c,k+1}=\begin{cases} 1, & \text{if } p\left(Conf_{t,p,c,k+1}\right) > \text{cut-off} \\ 0, & \text{otherwise} \end{cases}$$

where *cut-off* is a threshold lying within *(0, 1)*. In order to obtain the estimated probability *p(Conf)* machine learning techniques are applied to the available dataset including confirmable remarks actually observed in the previous reference dates; a cross-validation method allows to assess the cut-off level.

3. The NPL dataset and model selection

In this section we define the binary classification problem in order to predict whether a remark is confirmed by the RAs using the Italian NPL dataset.

The NPL dataset comprises detailed information on banking non-performing exposures and on the status of their credit recovery procedures transmitted by the parent company of a banking group that reports on a consolidated basis and by the individual banks on a stand-alone basis. Reporting follows a half-yearly periodicity with reference dates 31th December and 30th June (Banca d'Italia, 2016).

The dataset created for the study includes five reference dates from 30th June 2017 to 30th June 2019. In this time frame 445 RAs reported over 17 million of records assessed by a set of 37 automatic confirmable DQCs that generated 65,705 remarks.

The data used to estimate the probability of confirmation consider, as observations, all the individual confirmable remarks generated for each RA and reference date; more specifically, a remark is considered only once even if it is pending in more than one data submission. Remarks generated by non-confirmable DQCs are not used in the model estimation but are considered during the application of the decision-making rule.

As shown in Table 1, the total number of remarks detected in the given period amounts to 65,705, out of which 5,083 are 'confirmable remarks'; in turn, 4,643 of the confirmable remarks, related to the reference dates from 30th June 2017 to 31th December 2018, are used for the model estimation ('training set') and 440, related to the reference date 30th June 2019, for its assessment ('validation set').

Number of remarks — Table 1

Reference date	Number of remarks (total)	Number of confirmable remarks	Number of non-confirmable remarks
2017-06-30	31,306	758	30,548
2017-12-31	8,679	857	7,822
2018-06-30	18,706	1,398	17,308
2018-12-31	5,576	1,630	3,946
2019-06-30	1,438	440	998
Total	65,705	5,083	60,622

Sources: NPL dataset – Bank of Italy

The dataset contains the dummy variable for the observed confirmation of remarks (*Conf)* and the regressors. In particular, there are 430 dummy variables, of which 15 for DQCs and 415 for RAs, and numeric variables related to the imbalances defined as the differences among quantitative aggregates of the remark, the average number of records sent by an RA for a specific reference date, and the reference dates.

A wide set of commonly used supervised approaches were considered: ridge logistic classifier (Hoerl and Kennard, 1970; Le Cessie and Van Houwelingen, 1992; Schaefer *et al.*, 1984; Cule and De Iorio, 2012; Van Wieringen, 2020); linear and quadratic discriminant analysis (LDA and QDA); decision tree classifier; k-neighbors

classifier and random forest (James *et al.*, 2013; Hastie *et al.*, 2009). The results of the application of these approaches to the training set are reported in Table 2.

Model comparison in the training set and in the validation set — Table 2

	Model	Logistic regression	Ridge logistic classifier (λ=1)	Linear discriminant analysis	Decision tree classifier	Quadratic discriminant analysis	K-neighbors classifier	Random forest
	Optimal cut-off	0.41	0.69	0.50	0.52	0.49	0.53	0.71
Training set	Accuracy	0.8283	0.8277	0.7265	0.7252	0.7256	0.7271	0.5044
	Recall	0.9516	0.9168	0.9902	0.9355	0.9938	0.9807	0.3944
	Precision	0.8347	0.8558	0.7293	0.7483	0.7274	0.7330	0.8346
	Negative predictive value	0.7980	0.7303	0.5541	0.5023	0.5435	0.5390	0.3325
Validation set	Accuracy	0.8068	0.7841	0.7636	0.7523	0.7795	0.7545	0.7795
	Recall	0.9417	0.8455	0.9738	0.9038	0.9942	0.9359	1.0000
	Precision	0.8325	0.8735	0.7786	0.8031	0.7821	0.7887	0.7795
	Negative predictive value	0.6154	0.5093	0.1818	0.3889	0.5000	0.3333	NA

Sources: NPL dataset – Bank of Italy

The good performance of the logistic ridge regression with respect to other methods is not new in the literature (Bradley, 1996). Similar results have been obtained when the considered dataset had a large set of uncorrelated variables (Cornell-Farrow and Garrard, 2020), as it is the case of the NPL dataset. The logistic regression, as a special case of the ridge logistic classifier, has been selected for the model estimation since this method outperforms the others in terms of accuracy, recall and precision. The LDA, QDA and the random forest do not predict properly the probability of confirmation of a remark in the validation set, as (about) all the remarks are expected to be confirmed. QDA should exhibit a better performance in case of non-linear decision boundary, and in our empirical analysis its performance turned out to be only slightly higher than the LDA considering the validation set. A decision tree classifier and a k-neighbors classifier provide similar results than the previous models, but they still underperform the ridge logistic classifier. A few studies suggested that the latter models may improve their performance when filtering or clustering of features is applied (Ala'raj *et al.*, 2020; Rajaguru and Chakravarthy, 2019). However, this may be associated with enhanced skewness of the results, increasing their sensitivity and reducing their specificity.

Following Le Cessie and Van Houwelingen (1992) and considering that in our case the binary response *Conf* and a set of predictors X of U dimensions are measured on remarks, the estimated probability that a remark is confirmed is given by *p(Conf)*, where the probability function *p* follows the logistic regression model:

$$p(Conf) = \frac{exp(X\beta)}{1 + exp(X\beta)}$$

*T*he ridge logistic classifier is obtained by maximizing the likelihood function l(β) with a penalized parameter λ applied to all the coefficients β except the intercept (β_0); in such case the estimator will be:

$$\beta_{RL} = \underset{\beta}{\operatorname{argmax}} \left\{ l(\beta) - \lambda \sum_{u=1}^{U} \beta_u^2 \right\}$$

The logistic ridge regression estimator depends on the choice of a tuning parameter $\lambda \geq 0$ to be determined separately (Cule and De Iorio, 2013; Le Cessie and Van Houwelingen, 1992).

In order to maximize the model's performance, λ and the cut-off parameter were optimized by cross-validated grid-search. The best results are reached when the cut-off is 0.41 and λ is set to zero, i.e. when the standard logistic regression is applied.

4. Main results for the selected model

The results for the estimation of future confirmation of the remarks through the logistic regression model are shown in Section 4.1; the results related to the application of the proposed decision-making algorithm are presented Section 4.2.

4.1 Logistic regression: estimation of 'remark confirmed, Yes/No'

The logistic regression model used to estimate the probability that a remark is confirmed is defined as follows:

$$Logit(p(Conf)) = \beta_0 + \beta_1 \cdot Log(Imbalance) + \beta_2 \cdot Log^2(Imbalance) + \beta_3 \cdot Log(Records) +$$

$$+ \beta_4 \cdot Reference\_date + \sum_{c=1}^{15} \gamma_c \cdot DQC_c + \sum_{p=1}^{415} \delta_p \cdot RA_p + \varepsilon$$

The estimated confirmation $\widehat{Conf}$ of a remark, generated by a DQC c for the $(k+1)^{th}$ data submission sent by the RA p for the reference date t, is given by the following:

$$\widehat{Conf}_{t,p,c,k+1} = \begin{cases} 1, & if\ p\left(Conf_{t,p,c,k+1}\right) > 0.41 \\ 0, & otherwise \end{cases}$$

Where 0.41 is the optimal value for the cut-off since it maximizes the accuracy of the model. The model was estimated both on the training set and validation set. In general, all the measures computed in the training and validation sets suggest a satisfying performance of the chosen classification model (Table 3). The measures selected for assessing the goodness of the classification show high and similar values in the two sets, except for the negative predictive value that is 0.7980 and 0.6154, respectively, in the training set and in the validation set.

Since the purpose of the decision rule is to determine whether the DQL has improved by means of a subsequent data transmission, cases where remarks are predicted as confirmed when they should not (false positive) should be limited. This is a precautionary approach which however may lead to consider the DQL related to the new data submission as insufficient and request further analysis of remarks to the RA.

The results show that the model estimation achieves a high level of precision (0.8347 in the training set; 0.8325 in the validation set) and, at the same time, an

acceptable number of false negative cases (163 in the training set; 20 in the validation set).

Confusion matrix computed in the training set and the validation set — Table 3

Training set

		Conf		
		No	Yes	Total
$\widehat{Conf}$	No	644 (TN)	163 (FN)	807
	Yes	634 (FP)	3,202 (TP)	3,836
	Total	1,278	3,365	4,643

Measure	Value
Accuracy	0.8283
Recall	0.9516
Precision	0.8347
Negative Predictive Value	0.7980

Validation set

		Conf		
		No	Yes	Total
$\widehat{Conf}$	No	32 (TN)	20 (FN)	52
	Yes	65 (FP)	323 (TP)	388
	Total	97	343	440

Measure	Value
Accuracy	0.8068
Recall	0.9417
Precision	0.8325
Negative Predictive Value	0.6154

Note: TN true negative; FN false negative; TP true positive; FP false positive; *Conf* and $\widehat{Conf}$ are the actual and predicted values of the variable Conf, respectively.

Sources: NPL dataset – Bank of Italy

4.2 Application of the decision-making rule

As presented in Section 2, we analyse the application of the proposed decision rule to the 'case B' where two consecutive submissions are considered as releasable since they are not impaired by serious remarks. In particular, the decision-making algorithm (1) identifies whether a new report $k+1$ improves or worsens the DQL respect to the previous report k, sent by an RA p for the same reference date t.

An application on the NPL dataset has been performed by comparing in terms of DQL indicator all the 275 pairs of consecutive data submissions in the 'case B' with reference dates from 2017 to 2018 and the 14 ones referred to June 2019 (Table 4).

The results of the application of the decision rule confirm that the proposed method prevents the Data Manager from releasing non-fit-for-use data to the users since it automatically identifies, respectively for the reference dates 2017-2018 and June 2019, 29 and 1 cases where the DQL decreases with the subsequent submission (Table 5).

As expected the limited number of cases of DQL worsening is consistent with the RAs' data correction process that is typically characterised by an improvement of the

DQL over time. This consistency is further corroborated by the results related to a possible application of the decision rule to the other cases 'A', 'C' and 'D'[2].

Classification of submissions in the current approach

Table 4

Reference dates between years 2017 and 2018

		$(k+1)^{th}$ submission		
		Not-released	Released	Total
k^{th} submission	Not-released	269	407	696
	Released	51	275	326
	Total	320	682	1,002

Reference date June 2019

		$(k+1)^{th}$ submission		
		Not-released	Released	Total
k^{th} submission	Not-released	15	23	38
	Released	1	14	15
	Total	16	37	53

Sources: NPL dataset – Bank of Italy

Application of the decision-making algorithm to 'case B'

number and percentage in brackets

Table 5

Results of the decision rule	Reference dates between 2017 and 2018	Reference date of June 2019
Released submission	246 (89%)	13 (93%)
Additional Not-released submission	29 (11%)	1 (7%)
Total	275 (100%)	14 (100%)

Sources: NPL dataset – Bank of Italy

In order to assess the appropriateness of the classification of the submissions as released and additional not-released according to the decision rule, a comparison with a benchmark is carried out. In particular, the benchmark considered is obtained from the application of the decision rule when the observed variable *Conf* is considered instead of the estimated value according to the logistic model of Section 4.1.

$$\hat{I}_{t,p,k+1} \leq I_{t,p,k} \quad (2)$$

where $\hat{I}_{t,p,k+1}$ is computed using the estimated $\widehat{Conf}_{t,p,c,k+1}$.

The application of rule (1) is compared with (2) and the results are shown in Table 6 for the time period from 2017 to 2018 and June 2019, respectively. The results show that for the reference dates from 2017 to 2018, in 97% of cases the prediction

[2] See Appendix B from La Ganga *et al.*, 2022.

is effective; while, considering the reference date of June 2019, the decision is definitely the same in both cases.

Verification of the decision rule: comparison with the benchmark decision rule that considers the observed *Conf* Table 6

Reference dates between years 2017 and 2018

		Benchmark decision on the $(k+1)^{th}$ submission (using *Conf*)		
		Not-released	Released	Total
Decision on the $(k+1)^{th}$ submission considering $\widehat{Conf}$	Not-released	22	7	29
	Released	2	244	246
	Total	24	251	275

Accuracy = 0.9673

Reference date June 2019

		Benchmark decision on the $(k+1)^{th}$ submission (using *Conf*)		
		Not-released	Released	Total
Decision on the $(k+1)^{th}$ submission considering $\widehat{Conf}$	Not-released	1	0	1
	Released	0	13	13
	Total	1	13	14

Accuracy = 1

Sources: NPL dataset – Bank of Italy

Conclusions

In literature on data quality, this topic is generally approached by monitoring different dimensions, e.g. timeliness, accuracy and consistency of data. This paper proposes a decision-making rule that improves the monitoring of the accuracy, completeness and consistency of data sent by RAs to the Authorities and supports the Data Manager in deciding whether to immediately release the data to internal and external users.

More broadly, especially in all cases where data reporting is made more challenging for RAs due to exogenous and unpredictable events, such as also the pandemic fallout on organisational issues, the proposed DQL indicator can play an important role in making more efficient the DQM activity.

Using machine learning techniques based on the results of the automatic validation process and Data Manager's past evaluations of explanations received by RAs, we assess the variation of the DQL between two consecutive data submissions sent by the same RA and for a specific reference date. From a methodological point of view, a logistic regression model has been used to predict the confirmation probability of a single remark in order to estimate the difference in DQL between two versions of the same dataset (the second carrying the revisions of the first). The final model has been selected by comparing different available alternatives and optimizing the goodness of fit. In our case, the logistic regression model outperforms several models that are commonly used in machine learning studies.

In the second phase of the study, a decision-making rule has been developed by taking into account the estimated confirmation probability of a single remark, which emerged from the previous step, together with the total number and the severity of the remarks. The results of the application on the NPL dataset are remarkable, since in 97 per cent of cases the decision rule provides coherent results to those observed when the actual status is known.

In the current practice, the comparison between two consecutive reports submitted by the same RA is left to the judgment and expertise of the Data Manager and the interaction with the RA. The proposed decision-making rule leads to a less time-consuming and more harmonized approach. Moreover, it supports the Data Manager to determine whether data revisions do improve the DQL and it provides guidance for prioritizing DQM activities. Although the methodology is applied to a specific (granular) dataset – the banks' Non-Performing Loans reporting of the Bank of Italy – the method can be applied to other datasets in order to improve the DQL assessment, provided that past evidence from the DQM activity consisting of outliers, their severity and the confirmation is available.

References

Ala'raj, M., Majdalawieh, M., Abbod, M. F. (2020), 'Improving binary classification using filtering based on k-NN proximity graphs'. J Big Data 7, 15.

Banca d'Italia (2016), 'Instructions for the editing of reports on non-performing exposures'.

Bradley, A. P. (1996), 'The use of the area under the ROC curve in the evaluation of machine learning algorithms', The University of Queensland.

Casa, M., Graziani Palmieri, L., Mellone, L., Monacelli, F. (2022), 'The integrated approach adopted by Bank of Italy in the collection and production of credit and financial data', Occasional Papers series n. 667, Banca d'Italia.

Cornell-Farrow, S., Garrard, R. (2020), 'Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia, Communications in Statistics: Case Studies, Data Analysis and Applications', 6:2, 228-246.

Cule, E., De Iorio, M. (2012), 'A semi-automatic method to guide the choice of ridge parameter in ridge regression', Imperial College London and University College London.

Cule, E., De Iorio, M. (2013), 'Ridge regression in prediction problems: automatic choice of the ridge parameter', Genetic epidemiology, 37(7), 704-714.

Damia, V., Aguilar, C. P. (2006), 'Quantitative quality indicators for statistics an application to euro area balance of payment statistics'; Occasional Papers series n. 54, European Central Bank.

European Commission and Eurostat (2019), 'Quality Assurance Framework of the European Statistical System (V.2.0)'.

Gonzalez-Garcia, J. R., Pastor G. (2009), 'Benford's Law and Macroeconomic Data Quality' Volume 2009: Issue 010 International Monetary Fund.

Hastie, T., Tibshirani, R., Friedman, J. (2009), 'The elements of statistical learning: data mining, inference, and prediction', Springer Science & Business Media.

Hoerl, A., Kennard, R. (1970), 'Ridge Regression: Biased Estimation for Nonorthogonal Problems'. Technometrics, 12(1), 55-67.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), 'An introduction to statistical learning with applications in R', Springer Science & Business Media.

La Ganga, B., Cimbali, P., De Leonardis, M., Fiume, A., Meoli, L., Orlandi, M. (2022), 'A decision-making rule to detect insufficient data quality: an application of statistical learning techniques to the non-performing loans banking data', Occasional Papers series n. 666, Banca d'Italia.

Le Cessie, S., Van Houwelingen, J. (1992), 'Ridge Estimators in Logistic Regression'. Journal of the Royal Statistical Society, Series C (Applied Statistics), 41(1), 191-201.

Pipino, L. L., Lee, Y. W., Wang, R. Y. (2002), 'Data Quality Assessment', Communications of the ACM Vol. 45 (pages 211-218), No. 4ve.

Rajaguru, H., Chakravarthy, S. (2019), 'Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer', Asian Pacific journal of cancer prevention, 20(12):3777-3781.

Schaefer, R. L., Roi, L. D., Wolfe, R.A. (1984), 'A ridge logistic estimator, Communications in Statistics - Theory and Methods', 13:1, 99-113.

Schnabel, I. (2020), 'Don't take it for granted: the value of high-quality data and statistics for the ECB's policymaking', The ECB Blog, European Central Bank.

Tissot, B., de Beer, B. (2020), 'Implications of Covid-19 for official statistics: a central banking perspective', IFC Working Papers, n. 20, Bank for International Settlements.

Van Wieringen, W. N. (2020), 'Lecture notes on ridge regression' Vrije Universiteit Amsterdam.

BANCA D'ITALIA
EUROSISTEMA

A decision-making rule to detect insufficient data quality

an application of statistical learning techniques to the non-performing loans banking data

Barbara La Ganga, Paolo Cimbali, Marco De Leonardis, Alessio Fiume, Luciana Meoli and Marco Orlandi

Banca d'Italia

Motivation

- It is key to count on **an efficient and effective monitoring of the quality level of the data** transmitted by Reporting Agents (RAs) in order **to provide users with high-reliable data to carry out thorough and robust analyses.**

- **Data Quality Level (DQL) generally follows a positive trend** thanks to subsequent corrections submitted by RAs; however, **a data quality worsening may occur** especially when data production is affected by exogenous and unpredictable events, such as **RAs' IT malfunctions**, **changes in the reporting requirements** or **operative tensions and staff shortage** (also as seen during the pandemic).

- The aim of the study is to define a decision-making rule:
  - to speed up the **detection of DQL worsening;**
  - to provide a **synthetic measure of the DQL**.

Data quality cycle in Banca d'Italia

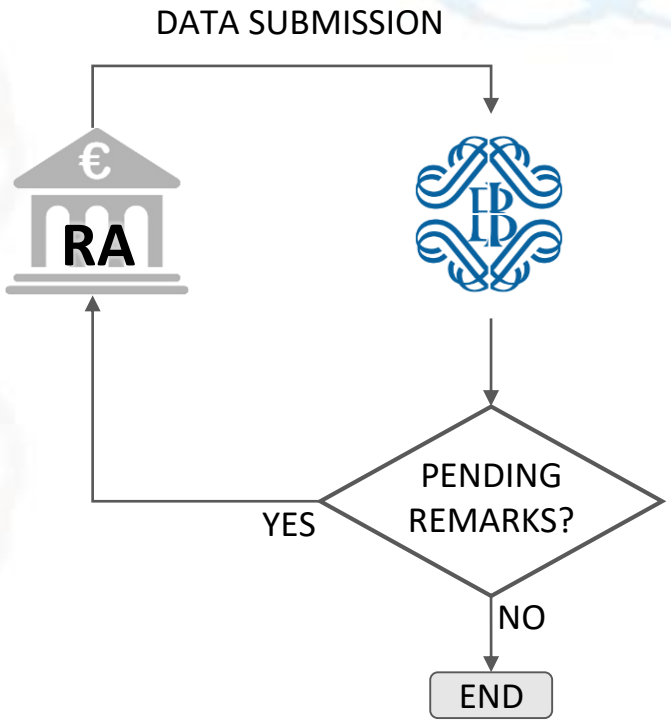
- At each data submission, the reliability of the data is assessed upon arrival by the Banca d'Italia by using a **set of automatic Data Quality Checks** (DQCs).

- A severity level from 0 to 10 is assigned to each DQC.

- When a DQC detects plausible errors (outliers) or deterministic errors, **remarks** are sent to the RA to request for:
  - **corrections** of erroneous data by sending a new data submission

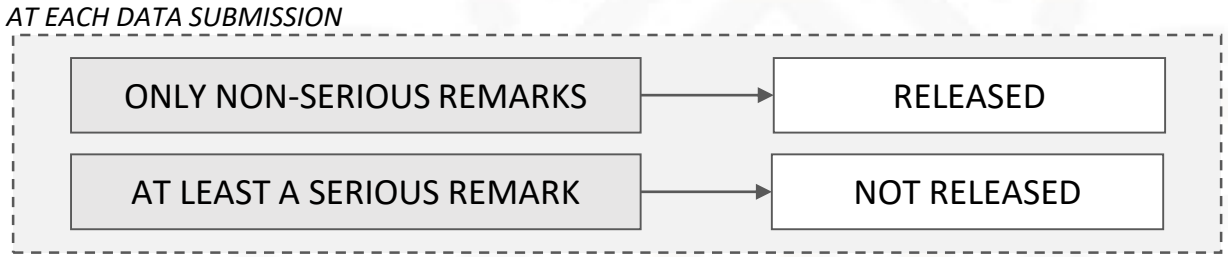
  or

  - **confirmations** of the data. These can be, in turn, accepted or refused by the Data Manager



Current decision rule to release data to users

- Based on the severity level of the DQC, the generated **remarks are classified as "serious" and "non-serious". If at least 1 serious remark is generated, the data submitted are kept on hold to be examined by the Data Manger** (hence not immediately released to the users)



- Considering 2 subsequent data submissions sent by an RA for a specific reference date, the possible cases are as follows:

		$(k+1)^{th}$ submission	
		Not-released	Released
k^{th} submission	Not-released	**D**	**C**
	Released	**A**	**B**

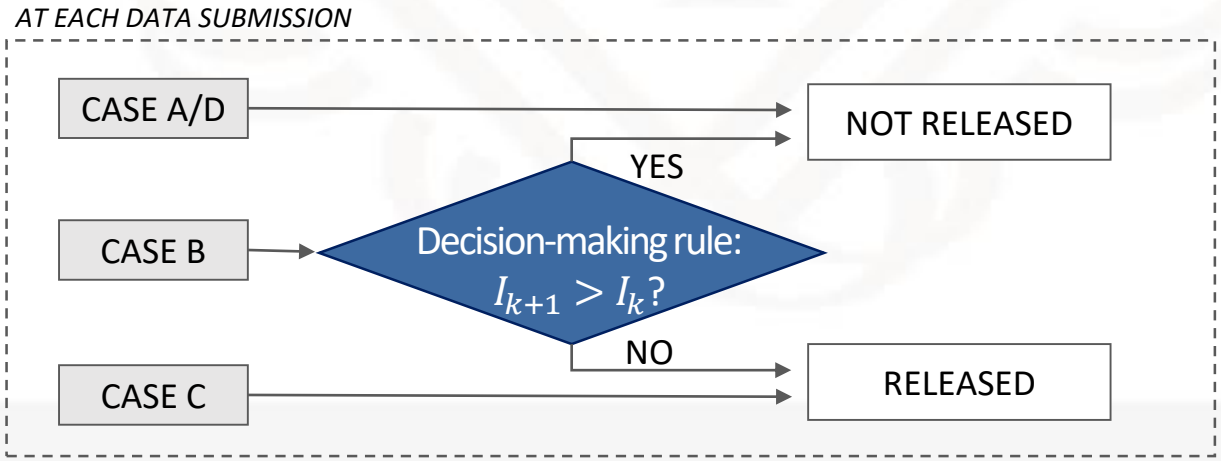
- **In cases A, C and D**, the Data Manager's decision is **straightforward;** in case B the $(k+1)^{th}$ submission may worsen the DQL.
- **The proposed decision-making rule is applied to case B** to detect the unexpected worsening of the DQL.

Definition of the proposed decision-making rule

- The proposed rule is based on a synthetic data quality indicator computed through past evidence from the Data Quality Management (DQM) activity:
  - **number of remarks** (**R**) generated by the **DQC *c***
  - **severity level** ($\boldsymbol{\tau}$)
  - **number of confirmations** (**Conf**)
- Definition of **a synthetic data quality indicator $\mathbf{I_k}$** for the k^{th} data submission sent by an RA for a specific reference date:

$$I_k = \sum_c \tau_c \cdot \left(R_{c,k} - Conf_{c,k}\right)$$

- If the DQC detects deterministic errors precisely (non-confirmable DQCs), $Conf_{c,k}$ is by construction equal to 0.
- **The higher the value of $\mathbf{I_k}$, the lower the DQL** of the k^{th} data submission.
- The proposed decision-making rule is defined as follows:



What quantities are available for the calculation of I_k and I_{k+1}?

- Let us assume we want to compare the DQL of the *(k+1)th* data submission with the DQL of the *kth*. Once the *(k+1)th* data submission is received, the **availability of the information for the calculation of I_k and I_{k+1}** is as follows:

	I_k	I_{k+1}
Number of remarks	✓	✓
severity level	✓	✓
Number of confirmations	✓	✘

- The **number of confirmations related to remarks, generated by the confirmable DQC c for the *(k+1)th* data submission, is estimated**:

$$\widehat{\text{Conf}}_{c,k+1} = \sum_{r=1}^{R_{c,k+1}} \widehat{\text{Conf}}_{c,k+1,\,r} \qquad where: \quad \widehat{\text{Conf}}_{c,k+1,\,r} = \begin{cases} 1, & if\ p\left(Conf_{c,k+1,\,r}\right) > cut-off \\ 0, & otherwise \end{cases}$$

- ***cut-off*** is a threshold lying within (0, 1) assessed with a cross-validation method
- The **estimation of the probability *p(Conf)*** is derived applying **machine learning techniques** to a dataset including remarks generated by confirmable DQCs actually observed in the previous reference dates.

Dataset and Model selection

- **Dataset: Banks Non-performing loans dataset (NPL)**, collected by Banca d'Italia on a biannual basis
  - over 17 million of records between 30*th* June 2017 and 30*th* June 2019
  - about 65K remarks generated, of which 5,083 by confirmable DQCs
  - 15 dummy variables for DQCs and 415 for RAs
  - numeric variables: differences among quantitative aggregates of remarks, number of records sent and reference dates
- **Model selection: the logistic regression model outperforms.**

	Model	Logistic regression	Ridge logistic classifier (λ=1)	Linear discriminant analysis	Decision tree classifier	Quadratic discriminant analysis	K-neighbors classifier	Random forest
	Optimal cut-off	*0.41*	*0.69*	*0.50*	*0.52*	*0.49*	*0.53*	*0.71*
Training set • *from June 2017 to December 2018* • *4,643 remarks*	**Accuracy**	0.83	0.83	0.73	0.73	0.73	0.73	0.50
	Recall	0.95	0.92	0.99	0.94	0.99	0.98	0.39
	Precision	0.83	0.86	0.73	0.75	0.73	0.73	0.83
	Negative predictive value	0.80	0.73	0.55	0.50	0.54	0.54	0.33
Validation set • *June 2019* • *440 remarks*	**Accuracy**	0.81	0.78	0.76	0.75	0.78	0.75	0.78
	Recall	0.94	0.85	0.97	0.90	0.99	0.94	1.00
	Precision	0.83	0.87	0.78	0.80	0.78	0.79	0.78
	Negative predictive value	0.62	0.51	0.18	0.39	0.50	0.33	NA

Sources: NPL dataset – Banca d'Italia

Application of the decision-making rule

- Considering the subsequent submissions of the case B, the decision-making rule allows the Data Manager to **automatically and promptly identify cases where the DQL decreases** and it **prevents the users to use non-fit-for-use data**.

Reference dates between years 2017 and 2018

		(k+1)th submission		
		Not-released	Released	Total
kth submission	Not-released	269	407	696
	Released	51	275 (Case B)	326
	Total	320	682	1,002

Results of the decision rule for the *(k+1)th* submissions of Case B	Percentage
Released submissions	89%
Additional Not-released submissions	11%

Reference date of June 2019

		(k+1)th submission		
		Not-released	Released	Total
kth submission	Not-released	15	23	38
	Released	1	14 (Case B)	15
	Total	16	37	53

Results of the decision rule for the *(k+1)th* submissions of Case B	Percentage
Released submissions	93%
Additional Not-released submissions	7%

Sources: NPL dataset – Banca d'Italia

Conclusions

- The proposed decision-making rule **improves the current DQL monitoring** by promptly detecting additional cases of DQL worsening.
- The synthetic data quality indicator I_k provide a **synthetic measure of the overall quality of data** transmitted by the RAs.
- **The decision-making rule is accurate**. It was assessed by comparing its results with the outcome resulting from an application of the decision-making rule based on the real status of the remarks confirmability: in 97% of cases the conclusions coincide.
- The proposed method can be **flexibly applicable to various data collections**.
- For the NPL dataset, the **implementation of the decision-making rule** in the Banca d'Italia's collection system is **ongoing**.

BANCA D'ITALIA
EUROSISTEMA

Thank you for your attention!

barbara.laganga@bancaditalia.it