IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Novel methodologies for data quality management
Anomaly detection in the Portuguese central credit register[1]

André Faria da Costa, Francisco Fonseca and Susana Maurício,
Bank of Portugal

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

Novel Methodologies for Data Quality Management

Anomaly Detection in the Portuguese Central Credit Register[1][2]

André Faria da Costa

Banco de Portugal, Lisboa, Portugal – anfcosta@bportugal.pt

Francisco Fonseca

Banco de Portugal, Lisboa, Portugal – ffonseca@bportugal.pt

Susana Maurício

Banco de Portugal, Lisboa, Portugal – scmauricio@bportugal.pt

Abstract

Since 2018, with the launch of the new Central Credit Register (CCR) covering, on a loan-by-loan basis, all loans to legal and natural persons and the corresponding credit risk, the granularity and volume of data collected by Banco de Portugal in this context have increased tremendously. The CCR receives roughly 20 million highly granular records every month, spread across more than 200 attributes, corresponding to individual instruments. The collection of such granular data has led to new challenges concerning data quality management (DQM), in particular how to detect in an efficient and effective way possible outliers in particular instruments or in a specific attribute of a particular instrument.

Due to the complexity of the CCR database, a single model would likely be unable to detect the majority of potentially anomalous data. In this paper, we will showcase the implementation of a set of methodologies which seek to cover as much potentially anomalous data as possible. The starting point of this process was the exploration of the Isolation Forest algorithm, which is based upon random forests and specifically designed for anomaly detection in large data sets. Afterwards, we have developed additional methodologies which complement this algorithm. Finally, we have implemented a Power BI dashboard where the results are presented, taking advantage of its visuals. This approach has allowed for the integration of all the different outputs from the DQM processes, and it notoriously increased the usability of results by the analysts.

Keywords: Banco de Portugal; Central Credit Register; Data Quality Management; Outlier Detection; Anomaly Detection; Isolation Forest

JEL classification: C80

[1] The views expressed are those of the authors and do not necessarily reflect those of Banco de Portugal.

[2] Acknowledgements: We would like to express our gratitude to Homero Gonçalves and Marta Veloso for all their valuable contributions and insights during this project and also to António Simões for helping us set up the IT Infrastructure. We would also like to thank André Fernandes, Daniel Peixeiro, Pedro Cordeiro, Rita Pisco and Rui Purificação for their collaboration in the team that provided the foundation for the development of these new tools. And last, but not least, we are thankful to Mário Lourenço for his revision and valuable suggestions.

Contents

1. Introduction

The use of machine-learning (ML henceforth) based methods for the purpose of anomaly detection has gained significant traction since the beginning of the 21st century. This increase in popularity is in part due to technological advancements in the field of ML as a whole, but also due to a massive increase in data granularity and complexity. Faced with massive and complex databases, more traditional outlier detection methods will (in general) perform poorly, especially from the computational point of view.

In the context of this paradigm shift regarding data volume and complexity, Banco de Portugal is not an exception. This is particularly apparent since the launch of the new Central Credit Register (CCR henceforth) system, in September 2018, which operates on an instrument-by-instrument logic, rather than a debtor-by-debtor one. Additionally, this new system follows a single service desk approach, which resulted in a widening of the set of variables reported. These changes resulted in a huge increase in the volume of information to be analysed - over 20 million monthly records, each characterized by more than 200 possible attributes.

Furthermore, as a central bank, data quality is of the utmost importance for Banco de Portugal. As such, the development and implementation of new data quality control methodologies, ML-based or otherwise, that support the management of this large and complex database becomes a necessity.

To this end, in late 2019, a team at the Statistics Department began researching novel methodologies that could potentially be used for data quality control on the CCR database, some of which were ML based outlier detection methods. The main objective of the team was the development and implementation of methodologies which were effective at identifying anomalies, with a low false negative rate and a reasonable false positive rate, while also being efficient from a computational point of view. Furthermore, our data is unlabelled, and for this reason, we looked mostly at unsupervised learning ML algorithms.

In this paper, we will highlight the three methodologies that were put into production as a result of the research conducted by this team:

- A pattern based test, which aims to detect potential inconsistencies in the reporting of any given instrument|debtor pairing in the database;
- A concentration check to oversee the evolution of the reporting of the categorical variables that are harder to handle for most outlier detection algorithms, including those based on ML models;
- An application of the Isolation Forest (IF henceforth) algorithm, which as the name indicates is a type of random forest based algorithm specifically designed for the purpose of isolating potentially anomalous observations. In our case, we use the IF algorithm to identify potentially anomalous behaviour in the evolution of reported outstanding amounts.

The remainder of this paper is organized in the following way: in section 2, we briefly describe the Portuguese CCR. In section 3, we discuss the implemented methodologies in detail. In section 4, we describe the way each test was implemented and the way the results are presented to the analysts. In section 5, we present our main conclusions and discuss the effectiveness of these algorithms across the full year that they have been used in production.

2. The Portuguese Central Credit Register

The Portuguese CCR is a system managed by Banco de Portugal, which gathers on a monthly basis a wide range of information provided by the observed agents (credit-granting institutions) associated with actual and potential credit liabilities of their costumers (natural or legal persons).

The main purpose of the CCR is to provide support to the credit-granting institutions in their assessment of counterparty risk. The registered entities have access to all the credit liabilities of their (actual and potential) customers, vis-à-vis the entirety of the resident financial system.

The data reported to the CCR, which is rich in both volume and complexity, is used for a variety of purposes – compiling statistics, banking supervision, financial stability analysis, informing monetary policy decisions, among many other possible uses.

Due to the multitude of purposes for which CCR data is used, data quality assurance is of the utmost importance. Prior to the present work, there were already multiple tests in place to ensure the consistency and coherence of the reported information, including:

- A set of validation tests that ensure that the reporting institutions abide by the established reporting rules;
- Cross-validation with other (internal and external) data sources that have some overlap with information reported to the CCR (like data from the Central Balance Sheet database, from Securities Statistics and from Statistics Portugal, among others);
- Analysis of the most significant (absolute and relative) month-on-month and homologous variations in the reported amounts, with the intent of identifying potentially anomalous variations that may be an indication of reporting errors.

As was previously stated, the CCR is large both in volume, as can be seen in Figure 1, but also in complexity – each record is characterized by more than 200 possible attributes, both quantitative (for example, outstanding amounts) and categorical (for example, the purpose of the credit). For this reason, and due to the importance of the CCR as a data source for both internal and external users, the need for new methodologies that complement the above-mentioned tests became evident. In the next sections, we will describe some of the methodologies that were developed and implemented with the intent of increasing the efficiency and the effectiveness of the data quality assessment process.

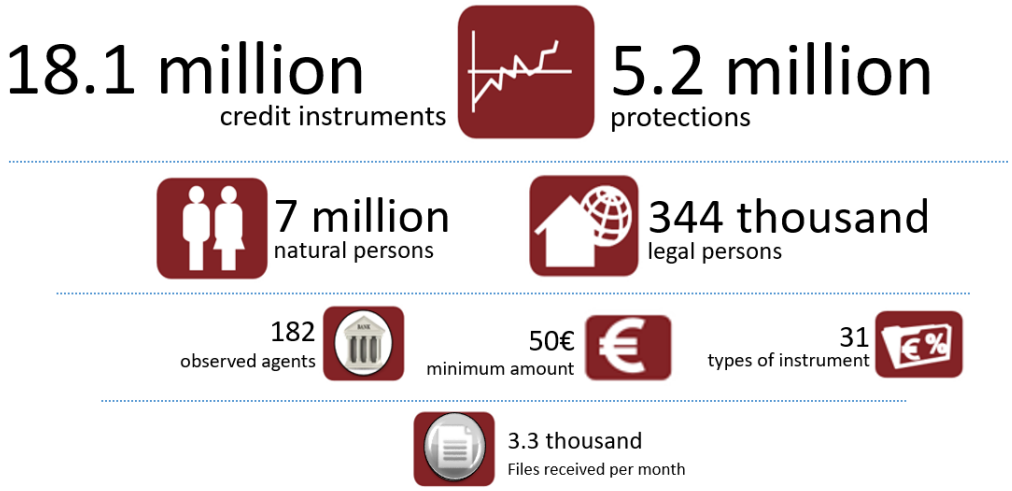


Figure 1 CCR indicators infographic

3. Selected Methodologies

Reporting Consistency Test

As mentioned in the previous section, the CCR receives information on a monthly basis, and as such it is highly relevant to ensure that each instrument is reported consistently until maturity. To this end, we developed a pattern-based test with the purpose of evaluating the reporting stability on an instrument-by-instrument and debtor-by-debtor basis.

The test looks at the last six months of a given instrument|debtor's lifetime and assigns a number between 0 and 2 based on the instrument|debtor's status in a given month:

- A value of 0 indicates that the instrument|debtor is not present in the database;
- A value of 1 indicates that the instrument|debtor is present in the database as active;
- A value of 2 indicates that the instrument|debtor is present in the database as finalised.

Taking into account the very large volume of information associated with this test (6 months of information, with over 20 million instrument|debtor pairs each month), we further define five types of patterns that are likely a sign of a reporting error, and as such should be analysed further:

- Instruments that are reported as finalised with non-zero outstanding amounts;
- Instruments reported as finalised and then as active in a following period;
- Instruments with reporting gaps (for example reported at time n, not reported at time n+1, and reported at time n+2);
- Instruments that cease to be reported without being finalised;
- Instruments that are reported as finalised for multiple periods (an instrument should be reported as finalised only once, and then it should not be reported in subsequent months).

Furthermore, to provide further context when presenting the results of this test, we also present the corresponding amounts. This allows us to quantify the impact that these potentially anomalous instruments have on the stock of credit as a whole.

Concentration Check

The concentration check test was developed with the purpose of tackling categorical variables that are reported to the CCR, since they are generally harder to handle for most anomaly detection algorithms.

For each combination of variable/observed agent/type of instrument, and for the latest four months of information reported, we compute the percentage of instruments that are reported with each possible element for the categorical variable. We also compute the percentage when considering the totality of instruments reported to the CCR (global averages).

As an example, if we are analysing the variable "Type of negotiation", for which the possible elements are "New operation", "Automatic renewal", "Regular renegotiation" and "Renegotiation due to default", we will compute the percentage of instruments reported with each of these possible elements, for both every individual institution and for the totality of the system, in the last four months.

The purpose of this test is two-fold:

- We can see if, for a given variable, the weight of any given element, within a given observed agent, has changed significantly in the last four months;
- We can see if, for a given variable, the weight of a given element, within a given observed agent, differs significantly from the weight when considering the totality of the system.

Furthermore, it is relevant to note that some of the variables which are utilized in the IF model are categorical in nature, meaning that this test also helps to ensure the quality of the data that is fed into the IF algorithm.

Isolation Forest

In order to select the ML model to be implemented, we had to take into account that, due to the volume and complexity of the database, and due to the way it was designed, we do not have a variable that classifies a given observation as being correctly or incorrectly reported. Hence, the use of supervised methods was excluded *a priori*, since we do not have a target variable.

From the pool of unsupervised methods, we considered two main model branches – clustering models (in particular, DBSCAN) and density-based models (in particular the IF algorithm).

In the case of clustering models, the DBSCAN algorithm performed well in testing. However, its complexity is $O(n^2)$, and since we are dealing with a very large dataset, the algorithm didn't perform well from a computational standpoint. This is particularly relevant when taking into consideration the fact that our intention was for the model to be ran frequently. Furthermore, estimating parameters is challenging, since DBSCAN is quite sensitive to small changes in its parameters.

Concerning density-based models, we tested the IF algorithm and found that it had many characteristics that we found advantageous:

- It performed well in testing;
- It has $O(n \log n)$ complexity (quasilinear);
- It is a scoring model, allowing us to establish a priority list that will let the analysts focus on the observations which have a higher likelihood of being anomalous;
- The task of parameter estimation is simpler than in the case of DBSCAN, since the algorithm is much less sensitive to small changes in its parameters.

This algorithm was initially proposed and described in detail in (Liu, Ting, & Zhou, 2009).

Summarily, the IF algorithm identifies anomalies through a process the authors refer to as a process of isolation – we select a subsample of our dataset and perform successive random partitioning, by first randomly selecting a variable and then randomly selecting a split value between the maximum and minimum values for that variable.

This process of recursive partitioning results in a binary tree structure. A node in the tree is considered a terminal node if it contains a single observation, or if all the observations contained within it have the same values for all attributes.

The partitioning process is carried out recursively until all nodes are terminal nodes. To improve performance, the process can also be stopped prematurely when we reach a pre-defined maximal tree depth.

Rather intuitively, the smaller the number of partitions needed for an observation to end up in a terminal node, the more isolated it is, and hence the likelier it is to be an anomaly. The model thus attributes a higher anomaly score to the observations that are more easily isolated.

Concerning model specification, it is relevant to establish that the main purpose of this model is the detection of unusual variations in the amounts reported to the CCR.

As was already mentioned in section 2, the CCR incorporates a wide variety of highly heterogeneous types of instrument; hence, we decided early on to calibrate a separate model for each type of instrument. Furthermore, we focused only on types of instrument that should have a regular payment, and for which that communication is mandatory.

Due to the nature of the IF algorithm, we also had to carefully choose the variables to be included – the model works through a process of isolation, so we had to pick variables where values (or combinations of values) that deviate from the norm are more than likely anomalies. For this reason, rather than using the variations of the raw reported amounts as model variables, we combined a set of variations that, generally, should cancel each other out into a new variable (henceforth referred to as "Residual Variation"):

$$\text{Residual Variation} = \text{Abs}(\Delta\text{Outstanding nominal amount} + \Delta\text{Accumulated write-offs} + \text{Payment} + \Delta\text{Off-balance sheet amount} + \text{Early repayment})$$

The construction of this new variable not only ensures that large values will more than likely be an indication of an anomaly in the reporting, but also reduces the number of variables to be included in the model, as it combines five numerical variables, without significant loss of information. However, to provide further context when analysing the results, we also present the raw amount variations to the analysts.

Adding to this numerical variable, we also include a set of categorical variables that characterize the loan (numerically encoded due to high cardinality):

- Purpose of the credit;
- Type of negotiation;
- Residual maturity (in five year steps).

It is also relevant to note that we need to be concerned about the special case of instruments that have a payment frequency that is not monthly. For this kind of instrument, although a due payment is reported on a monthly basis, this payment will only be reflected in a variation of the outstanding amounts when it is due (once every three, six or twelve months, in general). Thus, in months where a payment is not due, we will have (simplifying) $Residual\,Variation = Payment$. To avoid these instruments from showing up as potential anomalies, we filter out instruments where a payment is reported but no variation is recorded in any of the remaining amounts in a given month. This significantly reduced the false positive rate.

We use five consecutive months of reported information, meaning we have four deltas (variations) – the first three are used to train the model, and the last one corresponds to the production month.

Finally, it should also be noted that, since our numerical variable is a variation, the model will not be able to evaluate instruments, which are reported for the first time in the production month (since we are unable to calculate variations).

4. Implementation and Main Results

Workflow and Technologies Used

After defining the methodologies we wished to implement, we had to outline an operational workflow and to settle on which tools were to be used in the implementation process.

The reporting consistency and concentration check tests were implemented through the use of SQL procedures, and the IF test was implemented in Python, where we use the IF algorithm implemented in the Scikit-learn library (Pedregosa, et al., 2011). The test results are stored in a SQL Server database, for all three tests. The use of SQL, whenever possible, is a natural consequence of the fact that all of the data that is reported to the CCR is stored in SQL databases. For this reason, using SQL (or, in the case of Python, connecting to it via ODBC) minimizes the time required to transfer large volumes of data, making the process as efficient as possible.

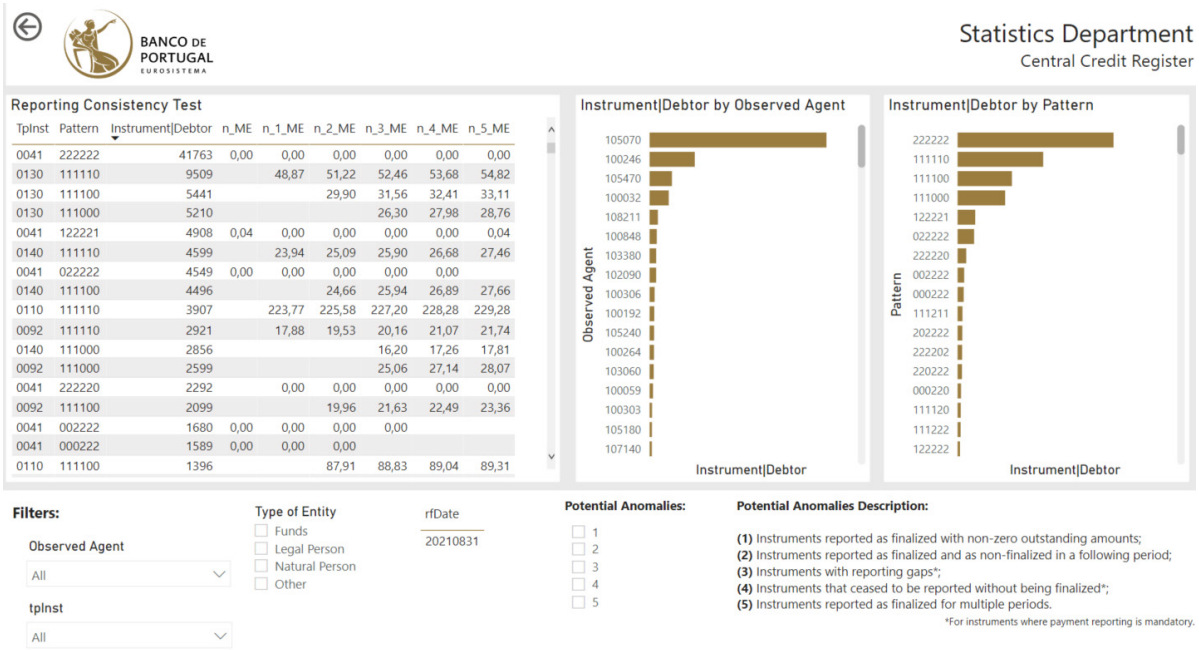
As a final step, for data exploration and visualization, we implemented a Power BI dashboard, that allows the analysts to easily access and interact with the results. The flexibility of the dashboard lets the analysts look at the results in the way that best suits the task at hand and select the cases that should be sent to the reporting institutions for further clarifications.

In the remainder of this section, we will provide real examples of anomalies that were detected and corrected through the use of each of the three tests we described previously. For confidentiality reasons, all identifying characteristics (observed agent, instrument and instrument identifiers) are anonymised.

Reporting Consistency Test

As described in the previous section, this test aims to evaluate the reporting consistency of the reporting institutions, through the construction of a pattern that characterizes the last six months of information for a given instrument|debtor pairing. The pattern is thus composed of six digits, where the last digit (rightmost) corresponds to the most recent month.

The results are made available in a Power BI dashboard, where we display the potentially anomalous patterns for each reporting institution and type of instrument. The use of Power BI makes the task of analysing the results more intuitive and far more flexible. Using filters, we can either look at the system as a whole, or select a single observed agent or type of instrument to narrow the scope of the analysis. The final dashboard (without any filtering) is shown in Figure 2.

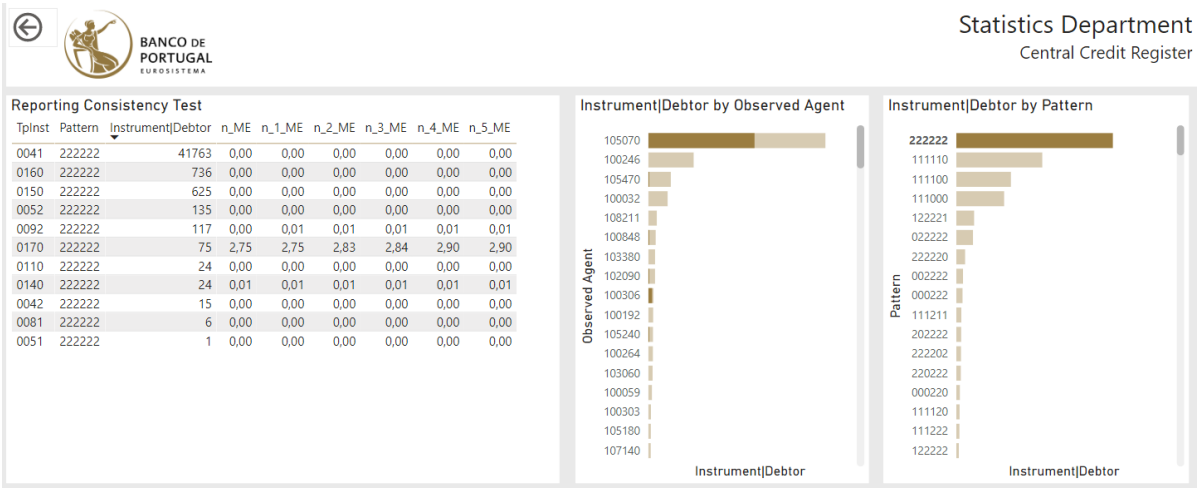


TpInst	Pattern	Instrument|Debtor	n_ME	n_1_ME	n_2_ME	n_3_ME	n_4_ME	n_5_ME
0041	222222	41763	0,00	0,00	0,00	0,00	0,00	0,00
0130	111110	9509		48,87	51,22	52,46	53,68	54,82
0130	111100	5441			29,90	31,56	32,41	33,11
0130	111000	5210				26,30	27,98	28,76
0041	122221	4908	0,04	0,00	0,00	0,00	0,00	0,04
0140	111110	4599		23,94	25,09	25,90	26,68	27,46
0041	022222	4549	0,00	0,00	0,00	0,00	0,00	
0140	111100	4496			24,66	25,94	26,89	27,66
0110	111110	3907		223,77	225,58	227,20	228,28	229,28
0092	111110	2921		17,88	19,53	20,16	21,07	21,74
0140	111000	2856				16,20	17,26	17,81
0092	111000	2599				25,06	27,14	28,07
0041	222220	2292		0,00	0,00	0,00	0,00	0,00
0092	111100	2099			19,96	21,63	22,49	23,36
0041	002222	1680	0,00	0,00	0,00	0,00		
0041	000222	1589	0,00	0,00	0,00			
0110	111100	1396			87,91	88,83	89,04	89,31

Figure 2 Dashboard view of the Reporting Consistency Test

On the right hand side, we can easily see the frequency with which each pattern shows up in our data, and the way these patterns are distributed amongst the observed agents. On the left hand side, we have a table where, for a given type of instrument and anomalous pattern, we show the number of instruments|debtors that display such a pattern, and their corresponding amounts in each of the six dates that make up the pattern. We can also perform a drill-down in order to drop to the level of each individual instrument|debtor, which often proves useful when contacting the reporting institutions to request clarification or correction.

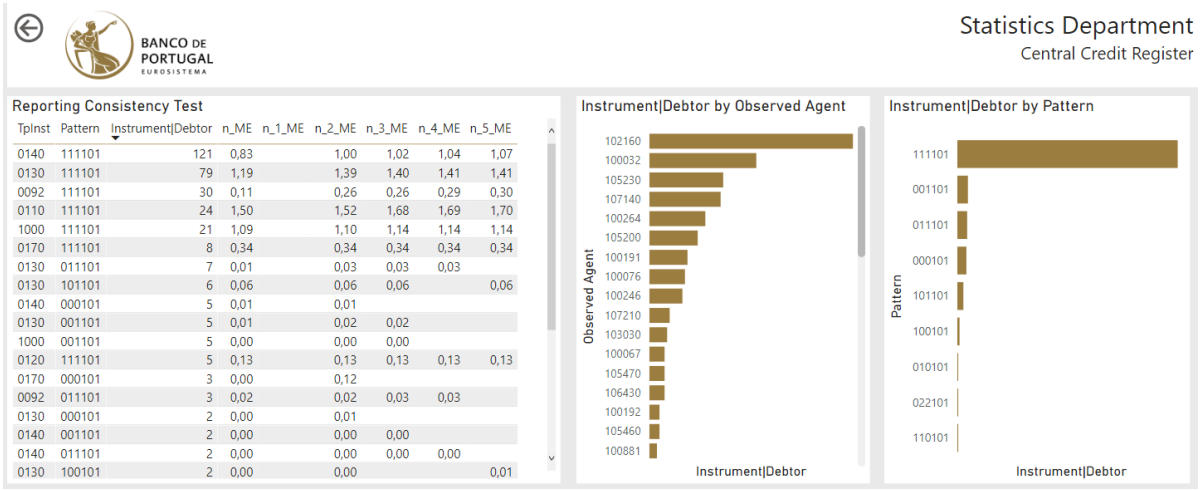
As an example of how the use of the presented filters can be helpful, if we filter for the pattern "222222" (Figure 3), we immediately see that institution 105070 is responsible for the vast majority of instruments that display this pattern. We can also see the corresponding number of instrument|debtor pairs and the amounts involved. In this case, as can be inferred in Figure 3, we have a few type 1 (Instruments reported as finalised with non-zero outstanding amounts) and multiple type 5 (Instruments reported as finalised for multiple periods) patterns. Even though we have detected just a few type 1 patterns, those cases are always analysed carefully since they indicate the existence of liabilities that are associated with terminated instruments, which is in general a sign of a reporting error with impact in the amounts.



TpInst	Pattern	Instrument|Debtor	n_ME	n_1_ME	n_2_ME	n_3_ME	n_4_ME	n_5_ME
0041	222222	41763	0,00	0,00	0,00	0,00	0,00	0,00
0160	222222	736	0,00	0,00	0,00	0,00	0,00	0,00
0150	222222	625	0,00	0,00	0,00	0,00	0,00	0,00
0052	222222	135	0,00	0,00	0,00	0,00	0,00	0,00
0092	222222	117	0,00	0,01	0,01	0,01	0,01	0,01
0170	222222	75	2,75	2,75	2,83	2,84	2,90	2,90
0110	222222	24	0,00	0,00	0,00	0,00	0,00	0,00
0140	222222	24	0,01	0,01	0,01	0,01	0,01	0,01
0042	222222	15	0,00	0,00	0,00	0,00	0,00	0,00
0081	222222	6	0,00	0,00	0,00	0,00	0,00	0,00
0051	222222	1	0,00	0,00	0,00	0,00	0,00	0,00

Figure 3 Reporting Consistency Test – filtering by a specific pattern

Type 3 patterns are also prioritised, since they show the existence of reporting gaps in the instrument|debtor's lifetime. As can be seen in Figure 4, we have a very small number of records that present a type 3 pattern and they are mostly associated with a gap at time n-1.



TpInst	Pattern	Instrument|Debtor	n_ME	n_1_ME	n_2_ME	n_3_ME	n_4_ME	n_5_ME
0140	111101	121	0,83	1,00	1,02	1,04	1,07	
0130	111101	79	1,19	1,39	1,40	1,41	1,41	
0092	111101	30	0,11	0,26	0,26	0,29	0,30	
0110	111101	24	1,50	1,52	1,68	1,69	1,70	
1000	111101	21	1,09	1,10	1,14	1,14	1,14	
0170	111101	8	0,34	0,34	0,34	0,34	0,34	
0130	011101	7	0,01	0,03	0,03	0,03		
0130	101101	6	0,06	0,06	0,06		0,06	
0140	000101	5	0,01	0,01				
0130	001101	5	0,01	0,02	0,02			
1000	001101	5	0,00	0,00	0,00			
0120	111101	5	0,13	0,13	0,13	0,13	0,13	
0170	000101	3	0,00	0,12				
0092	011101	3	0,02	0,02	0,03	0,03		
0130	000101	2	0,00	0,01				
0140	001101	2	0,00	0,00	0,00			
0140	011101	2	0,00	0,00	0,00	0,00		
0130	100101	2	0,00	0,00				0,01

Figure 4 Reporting Consistency Test – filtering by a specific anomaly type

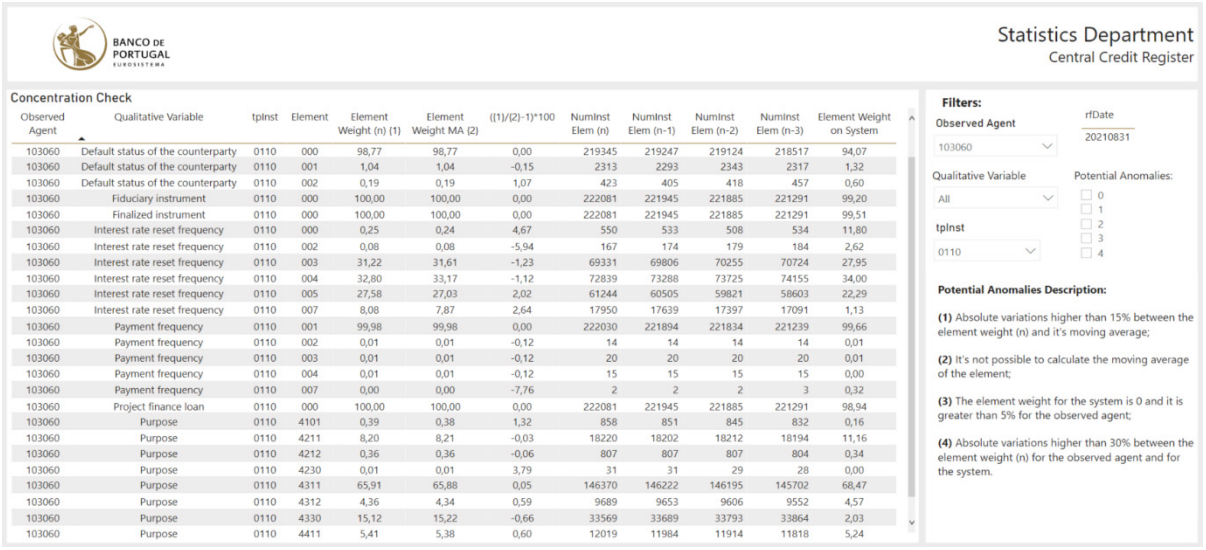
As was shown in the preceding examples, the set of potentially anomalous instrument|debtor pairs is quite low when taking into consideration the volume of the database. We can also see that, looking at the amounts involved, they are most significant in the case of type 4 patterns, that is, the observed agent ceased to report the instrument without first reporting it as finalised. Although this pattern may result from gaps in reporting, in most cases it simply means that the observed agent did not report the contract as finalised before ceasing to report it. Thus, this situation, while anomalous, was not our primary focus since it generally does not affect the conclusions we can derive from the CCR data, as finalised instruments do not have amounts different from 0.

Although the anomalies detected in this test may not result in a huge effect on the aggregates, they may have a significant impact for each individual debtor. Hence identifying them and requesting clarification or correction is highly relevant to the CCR's main purpose: provide the registered entities with information on all the credit liabilities of their (actual and potential) clients, vis-à-vis the resident financial system, in order to aid in their assessment of counterparty risk.

Concentration Check

As we have previously stated, the concentration check test allows us to evaluate the quality of the reporting of categorical variables to the CCR at a given moment, and it lets us assess how it has evolved over time. As with the remaining tests, this analysis is usually performed by looking at a single combination of observed agent and type of instrument, but we can also look at the state of the system as a whole. To aid in this analysis, we define four filters that are intended to highlight potential anomalies of different types.

As with the previous test, we display the results in a Power BI dashboard:



Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2})-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Default status of the counterparty	0110	000	98,77	98,77	0,00	219345	219247	219124	218517	94,07
103060	Default status of the counterparty	0110	001	1,04	1,04	-0,15	2313	2293	2343	2317	1,32
103060	Default status of the counterparty	0110	002	0,19	0,19	1,07	423	405	418	457	0,60
103060	Fiduciary instrument	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	99,20
103060	Finalized instrument	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	99,51
103060	Interest rate reset frequency	0110	000	0,25	0,24	4,67	550	533	508	534	11,80
103060	Interest rate reset frequency	0110	002	0,08	0,08	-5,94	167	174	179	184	2,62
103060	Interest rate reset frequency	0110	003	31,22	31,61	-1,23	69331	69806	70255	70724	27,95
103060	Interest rate reset frequency	0110	004	32,80	33,17	-1,12	72839	73288	73725	74155	34,00
103060	Interest rate reset frequency	0110	005	27,58	27,03	2,02	61244	60505	59821	58603	22,29
103060	Interest rate reset frequency	0110	007	8,08	7,87	2,64	17950	17639	17397	17091	1,13
103060	Payment frequency	0110	001	99,98	99,98	0,00	222030	221894	221834	221239	99,66
103060	Payment frequency	0110	002	0,01	0,01	-0,12	14	14	14	14	0,01
103060	Payment frequency	0110	003	0,01	0,01	-0,12	20	20	20	20	0,01
103060	Payment frequency	0110	004	0,01	0,01	-0,12	15	15	15	15	0,00
103060	Payment frequency	0110	007	0,00	0,00	-7,76	2	2	2	3	0,32
103060	Project finance loan	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	98,94
103060	Purpose	0110	4101	0,39	0,38	1,32	858	851	845	832	0,16
103060	Purpose	0110	4211	8,20	8,21	-0,03	18220	18202	18212	18194	11,16
103060	Purpose	0110	4212	0,36	0,36	-0,06	807	807	807	804	0,34
103060	Purpose	0110	4230	0,01	0,01	3,79	31	31	29	28	0,00
103060	Purpose	0110	4311	65,91	65,88	0,05	146370	146222	146195	145702	68,47
103060	Purpose	0110	4312	4,36	4,34	0,59	9689	9653	9606	9552	4,57
103060	Purpose	0110	4330	15,12	15,22	-0,66	33569	33689	33793	33864	2,03
103060	Purpose	0110	4411	5,41	5,38	0,60	12019	11984	11914	11818	5,24

Figure 5 Dashboard view of the Concentration Check

In Figure 5, on the right hand side, we have a set of filters that allow us to select both the observed agent (103060, in this case) and the type of instrument we want to focus on (0110, housing credit). In the case of this particular agent, we can see that for the displayed variables, the reporting is quite stable from either perspective:

- When looking at the way the agent's reporting evolves over time, by comparing the "Element Weight (n) {1}"[3] and "Element Weight MA {2}"[4] columns or looking at the variation (as displayed in the column labeled $\left(\frac{\{1\}}{\{2\}} - 1\right) * 100$), we find no significant changes;
- When looking at how the agent's reporting compares to the system as a whole, by comparing the "Element Weight (n)" and "Element Weight on System" columns, we see that the agent conforms to system trends.

The small number of potentially anomalous situations detected by this test, while also related to the fact that it deals with aggregate values (which in general are far more stable than individual records), is fundamentally a result of the arduous work carried out by the analysts at the Banco de Portugal CCR. Their work and subsequent interactions with the observed agents has resulted in numerous corrections since the test has been in production, which have improved the quality of the reported data very significantly.

In Figure 6 we can see an example of such a correction, concerning the agent/instrument pairing 103060/0110, which was corrected in August 2020.

[3] "Element Weight (n)" is the weight of the element on the reporting of the corresponding categorical variable, for a given observed agent, in the most recent reporting period. It is computed by dividing the number of contracts reported with the element in question (column "NumInst Elem (n)") by the total number of contracts reported by the agent for the type of instrument being considered.

[4] "Element Weight MA" is a weighted average of the three months prior to the current month, where the weights are 3/6 for period (n-1), 2/6 for period (n-2) and 1/6 for period (n-3). It is computed in the same way as "Element Weight (n)".

Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Purpose	0110	4101	0,34	0,10	227,31	755	236	228	219	0,06
103060	Purpose	0110	4211	8,06	1,07	651,88	18062	2440	2379	2325	6,16
103060	Purpose	0110	4212	0,36	0,05	612,31	815	116	114	110	0,21
103060	Purpose	0110	4230	0,01	0,00	285,44	27	7	7	7	0,00
103060	Purpose	0110	4311	65,66	15,28	329,65	147055	34203	34209	34182	72,77
103060	Purpose	0110	4312	4,11	0,58	610,30	9211	1312	1295	1249	2,61
103060	Purpose	0110	4330	15,68	0,25	6.060,53	35120	573	566	567	2,01
103060	Purpose	0110	4411	5,08	0,02	21.273,61	11372	52	54	55	4,01
103060	Purpose	0110	6000	0,69	82,63	-99,17	1543	184936	184918	184902	0,42

Figure 6 Concentration Check – purpose of loan (August 2020)

As we can observe, up to July 2020, this agent presented a concentration of over 80% in the element "6000 - Other purposes" for the variable "Purpose" (seen in column "Element Weight MA"), which is meant to be a residual element. This was corrected in the following month of August 2020, as can be seen in the "Element Weight (n)" column, and the instruments that presented this element were mostly reallocated across the remaining available elements, in particular many were moved to element "4311 - Residential real estate purchase – permanent residential property". The element weights displayed as of August 2020 for this variable also fell into line with the behaviour displayed by the system as a whole, which was not the case in July, as should be expected in general. Nevertheless, this correction triggered multiple type 1 anomalies alerting the analyst for the significant changes that took place. If we had instead focused on the picture as of July 2020, the very high concentration in the element "6000 - Other purposes", when compared with the system weights, would give rise instead to a type 4 anomaly.

After this correction the situation has remained quite stable, as can be seen in Figure 7 showing August 2021:

Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Purpose	0110	4101	0,39	0,38	1,32	858	851	845	832	0,16
103060	Purpose	0110	4211	8,20	8,21	-0,03	18220	18202	18212	18194	11,16
103060	Purpose	0110	4212	0,36	0,36	-0,06	807	807	807	804	0,34
103060	Purpose	0110	4230	0,01	0,01	3,79	31	31	29	28	0,00
103060	Purpose	0110	4311	65,91	65,88	0,05	146370	146222	146195	145702	68,47
103060	Purpose	0110	4312	4,36	4,34	0,59	9689	9653	9606	9552	4,57
103060	Purpose	0110	4330	15,12	15,22	-0,66	33569	33689	33793	33864	2,03
103060	Purpose	0110	4411	5,41	5,38	0,60	12019	11984	11914	11818	5,24
103060	Purpose	0110	6000	0,23	0,22	4,07	518	506	484	497	5,96

Figure 7 Concentration Check - purpose of loan (August 2021)

As a final example, in Figure 8, we present a recent case where a correction took place for the agent/type of instrument pair 103030/0130 (consumer credit), for the variable "Interest rate reset frequency".

Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103030	Interest rate reset frequency	0130	000	81,83			86847	1			78,76
103030	Interest rate reset frequency	0130	002	0,00	0,00	0,29	5	5	5	5	0,75
103030	Interest rate reset frequency	0130	003	0,35	0,36	-4,63	368	381	389	401	1,34
103030	Interest rate reset frequency	0130	004	15,09	15,65	-3,56	16014	16370	16772	17269	3,33
103030	Interest rate reset frequency	0130	005	1,69	1,74	-2,48	1798	1826	1862	1892	5,85
103030	Interest rate reset frequency	0130	007	1,04	82,25	-98,74	1104	87659	87483	87327	9,78

Figure 8 Concentration Check – interest rate reset frequency (August 2021)

This situation is quite similar to the previous one. The agent was reporting over 80% of instruments with element "007 – Other frequencies" (as can be seen in the "Element Weight MA" column). After

further investigation, we concluded that this issue was the result of a mapping error for the instruments with fixed interest rate. After contacting the reporting agent, a correction took place and the instruments that were previously classified with element "007" were reclassified as "000 - Rate cannot be reset" which is now the element with the most significant weight, in line with the system weights.

Isolation Forest

As was stated in the previous section, IF is a scoring algorithm, meaning that it provides us with a metric of how likely it is for a given observation to be an anomaly. This means we can easily provide the analysts with an ordered list of potential anomalies, according to the score attributed by the model.

As with the previous tests, the results are provided to the analysts using a Power BI dashboard, with a visual and an accompanying table, as displayed in Figure 9:

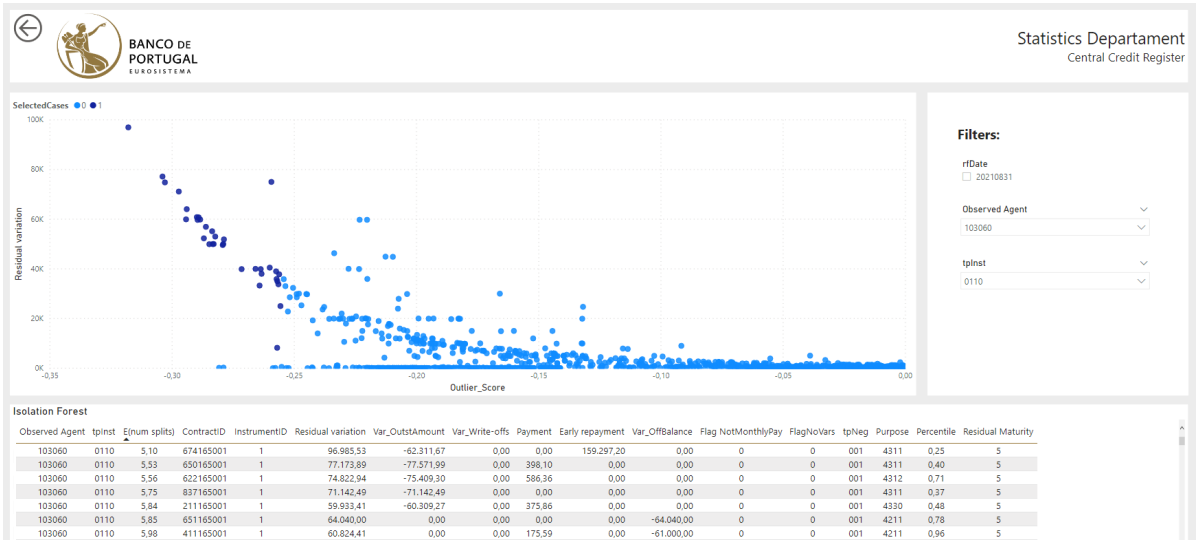


Figure 9 Dashboard view of the Isolation Forest results

In the accompanying table, along with the model variables ("residual variation", "purpose of the loan", "type of negotiation" and "residual maturity"), we also present a set of other variables to provide further context that could help the analysts determine if a given observation is, in fact, an anomaly:

- Identifiers of the observed agent and of the contract/instrument ("Observed Agent", "ContractID", "InstrumentID");
- Type of instrument ("tpInst");
- Expected value of the number of splits needed to isolate the data point in the IF ("E(num splits)");
- Variations of the amounts that, when combined, result in the value of the "residual variation";
- Flag that indicates if the instrument has a payment frequency which is not monthly ("Flag NotMonthlyPay");
- Flag that indicates if the instrument did not experience variations in any of the reported amounts (excluding the payment) in the current month ("FlagNoVars");
- Percentile of the outstanding nominal amount, considering all instruments reported for a given type of instrument, for a given observed agent, in the reference period.

In the visual, we show the distribution of the residual variations related to the outlier score, where the lower the score, the greater the probability of the observation being an anomaly.

All the observations that the test detects (regardless of anomaly score) are displayed in the visual. In dark blue, we differentiate the observations that should be prioritised when questioning the observed agents. An instrument is selected as being a priority if it meets all of the following criteria:

- Outlier score belongs to the 1st percentile of outlier scores, within the corresponding type of instrument/observed agent combination;
- Outlier score less than or equal to -0.25;
- Residual variation greater than or equal to €1,000.

As stated above, multiple situations can generate significant outlier scores, either due to the high value of the residual variation or due to the uncommon combination of the three categorical variables.

In Figure 10 we present a few examples where the residual variation is clearly the factor that determines the small "E(num splits)"[5]. The descriptions are presented in the order they are shown in the table:

Observed Agent	tpInst	E(num splits)	ContractID	InstrumentID	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance
105070	0110	5,05	045823733	1	130.100,00	0,00	0,00	130.100,00	0,00	0,00
103060	0110	5,10	674165001	1	96.985,53	-62.311,67	0,00	0,00	159.297,20	0,00
103060	0110	5,53	650165001	1	77.173,89	-77.571,99	0,00	398,10	0,00	0,00
103060	0110	5,85	651165001	1	64.040,00	0,00	0,00	0,00	0,00	-64.040,00
103060	0110	6,01	130165001	1	60.738,15	0,00	-60.738,15	0,00	0,00	0,00

Figure 10 Isolation Forest – quantitative variables impact

- Monthly payment is reported but the outstanding amounts do not decrease;
- Early payment is reported but the outstanding amounts do not decrease accordingly;
- The monthly payment significantly differs from the variation observed in the outstanding amounts;
- The off-balance sheet amount decreases but the outstanding amount does not increase;
- Decrease in the amount allocated to write-offs with no increase in the outstanding amounts.

In Figure 11, we highlight two data points that show that despite the fact that the residual variation seems to be the determining factor for the value of the outlier score, the categorical variables also have a significant effect. The highlighted data points have a significant outlier score, despite the fact that they have residual variations that are not very high.

[5] Expected value of the number of splits required to isolate a given observation. This value conveys the exact same information as the outlier score, i.e., the lower the number of splits required to isolate a data point, the higher the likelihood of the point being an anomaly.

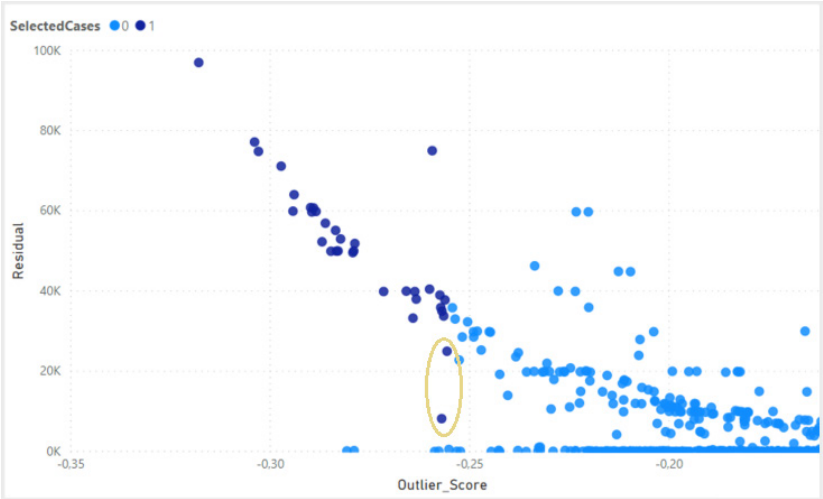


Figure 11 Isolation Forest – categorical variables impact (visual)

Furthermore, we can see that the data point with the lower "residual variation" actually has a more significant outlier score. If we analyse both data points in detail (Figure 12), we can see that this is because the data point with a higher outlier score has a "residual maturity" of -1 (passed maturity), which is likely an unusual element for this agent/instrument pairing, whereas the element 5 (Over 20 years) is far more common.

Outlier_Score	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance	tpNeg	Purpose	Percentile	Residual Maturity
-0,2570	8.202,70	-8.202,70	0,00	0,00	0,00	0,00	001	4311	0,02	-1
-0,2557	25.005,00	0,00	0,00	0,00	0,00	-25.005,00	001	4101	0,95	5

Figure 12 Isolation Forest – categorical variables impact (table)

5. Conclusion and Final Remarks

As we transited in September 2018 from a debtor-by-debtor logic to an instrument-by-instrument one with the new CCR, we experienced a huge increase in the volume of information to be processed and analysed. In addition, the new approach of a single service desk represented a widening in the set of variables related with the credit concession, with the addition of multiple new attributes associated.

This evolution represented a new challenge and it became of critical importance to find new ways of looking at data efficiently to detect and control for a multitude of anomalies arising from either the evolution or the interaction of the variables reported. This was the context that motivated the development of the three new tests we have presented: the reporting consistency test, concentration check and IF.

After a full year of usage in production, on a monthly basis, the new tests developed have shown to be very useful and we have received very encouraging feedback. They represent a valuable addition to the quality control process, focusing on dynamics that complement the other existing processes, allowing for the identification of a set of anomalies that previously would not be detected or would require complex and time-consuming *ad hoc* analyses. Examples of this kind of abnormal evolutions include:

- The detection of reporting gaps and strange patterns, even subtle ones that only affect a few instruments;
- Oversee the evolution of the reporting of categorical variables and to detect structural changes;
- Monitor the evolution of the amounts reported for the instrument, taking into account the categorical variables that characterize them and ranking them by the degree of severity for further questioning.

Hence, these new tools have contributed unquestionably to an increase in both the effectiveness and efficiency of the data quality assessment process and are an important enhancement to the analysts' tool set.

References

Liu, F. T., Ting, K., & Zhou, Z.-H. (2009). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, (pp. 413 - 422).

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.

BANCO DE PORTUGAL
EUROSISTEMA

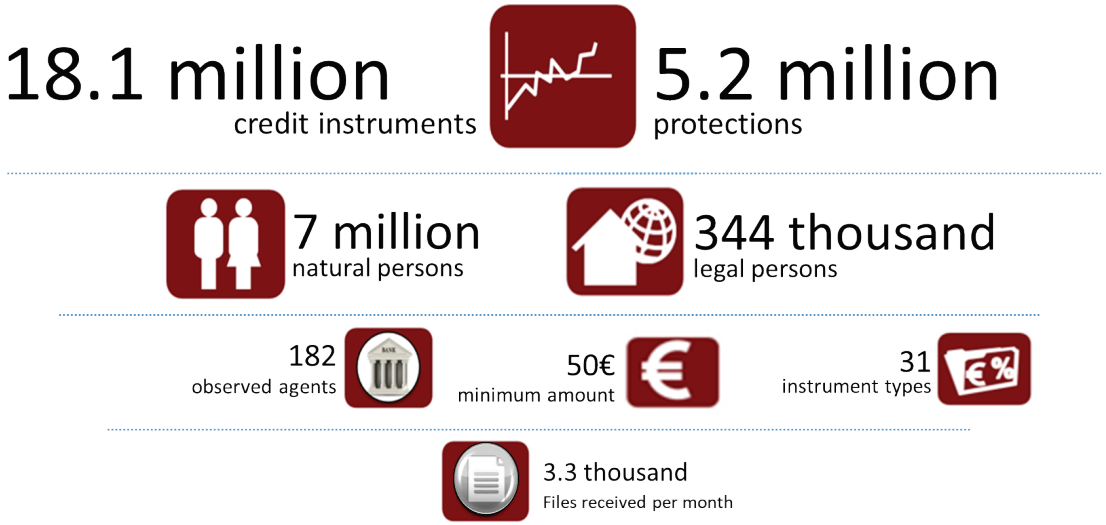
Anomaly Detection in the Portuguese Central Credit Register (CCR)

André Costa
Francisco Fonseca

22 October 2021

Introduction

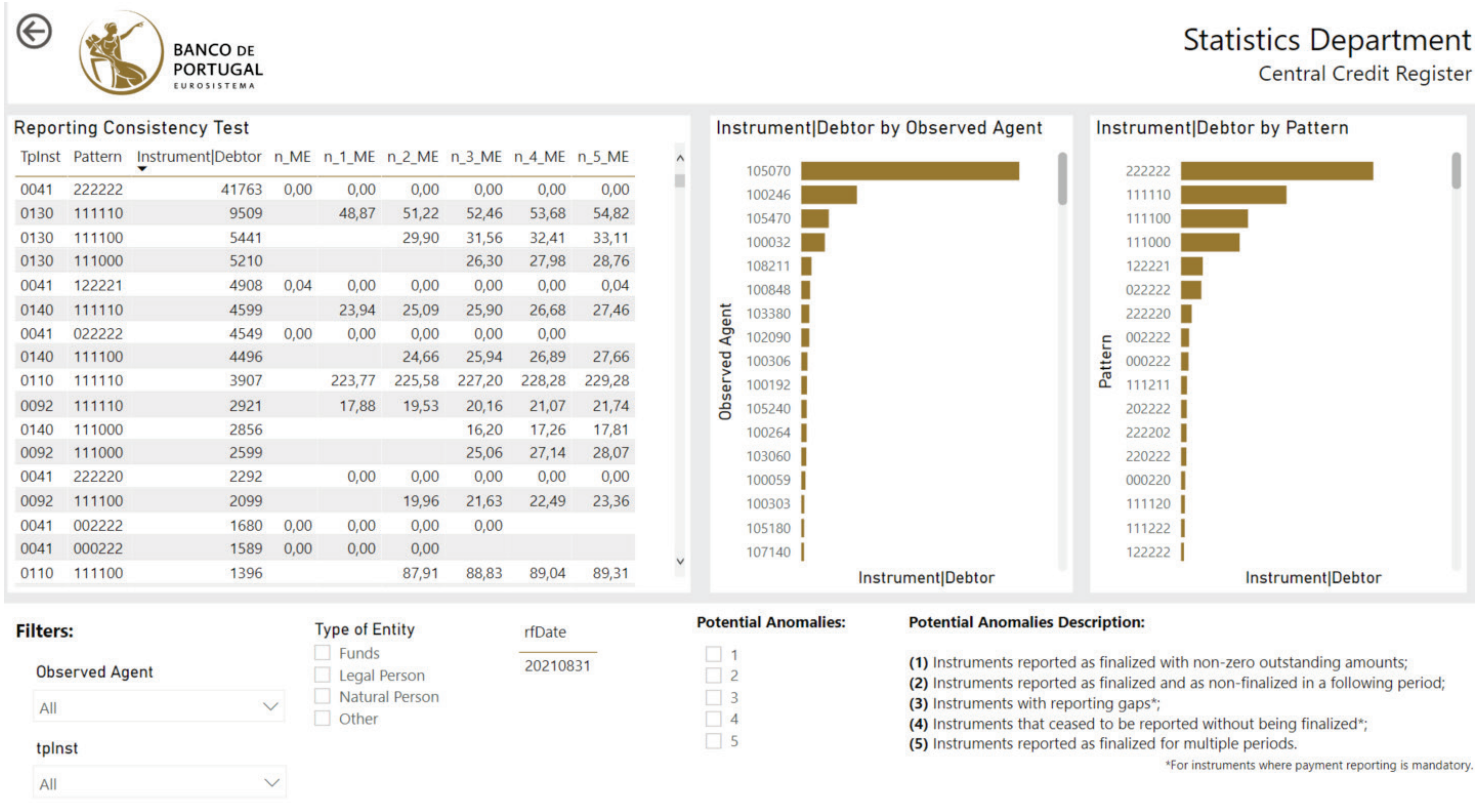
- The Portuguese CCR is a system which gathers on a monthly basis information associated with actual and potential credit liabilities of natural and legal persons.
- The main purpose is to provide support to the credit institutions in their assessment of counterparty credit risk.
- The data reported to the CCR, which is rich in both volume and complexity, is used for a multitude of different purposes – compiling statistics, banking supervision, financial stability analysis, support on monetary policy decisions…



- It’s quite challenging to perform the quality control of a database with such a level of granularity and variety of attributes (over 200). This was the main motivation that lead to the development of these new tests: increase the efficiency of the data quality controls, detect more subtle evolutions, create automatic filters for a range of potential anomalies and ensure the coherence of the process across the different observed agents.

Reporting Consistency Test

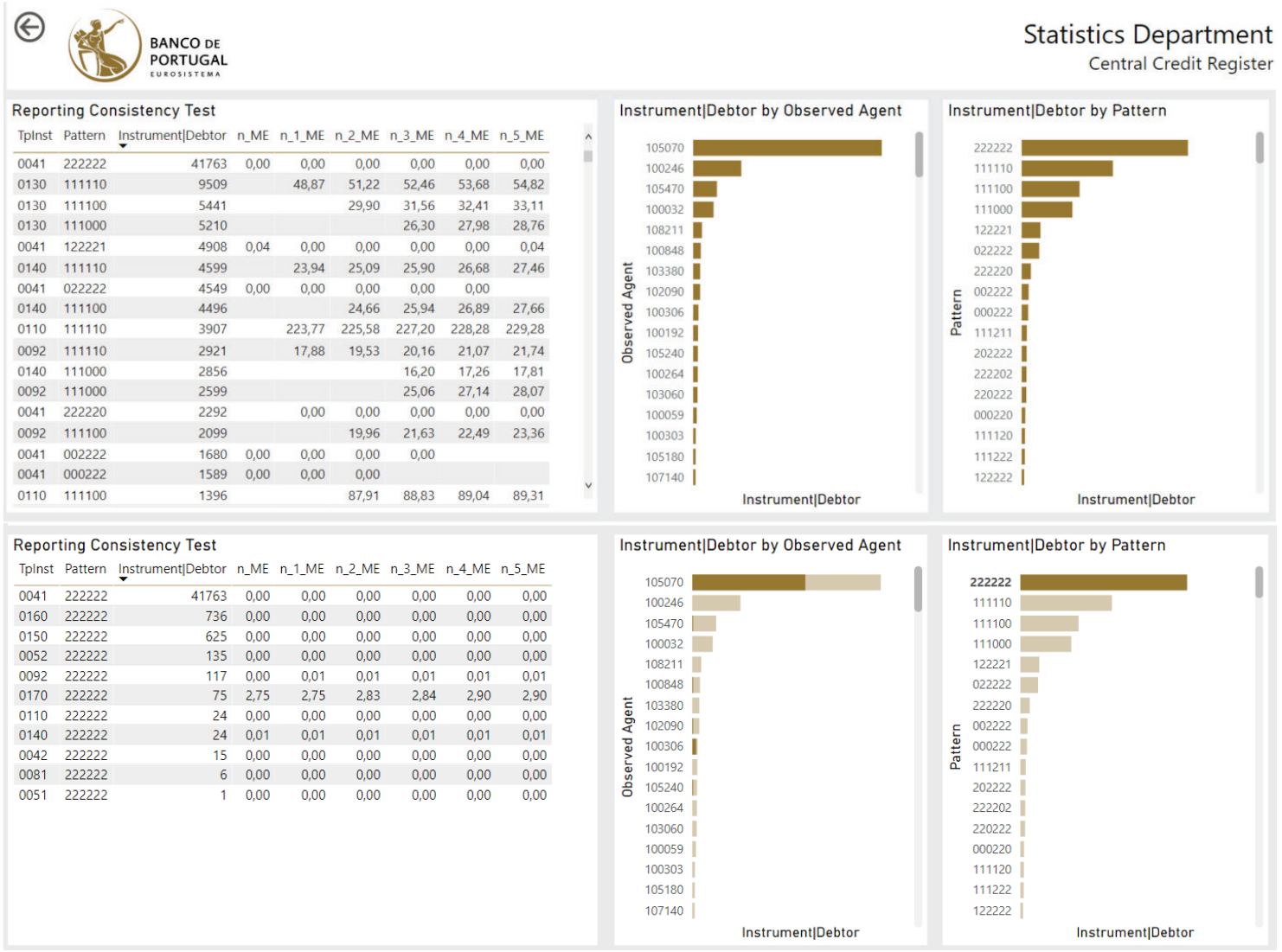
- The goal is to evaluate if all instruments are reported consistently until maturity;
- A pattern is created with the last six months of information reported (the rightmost digit is the most recent period);
- Pattern possible values for each month:
  - 0 – the instrument/entity was not reported;
  - 1 - the instrument/entity is active;
  - 2 – the instrument/entity was finalized.[1]



[1] Finalized instruments should be reported only once and with all amounts set to 0.

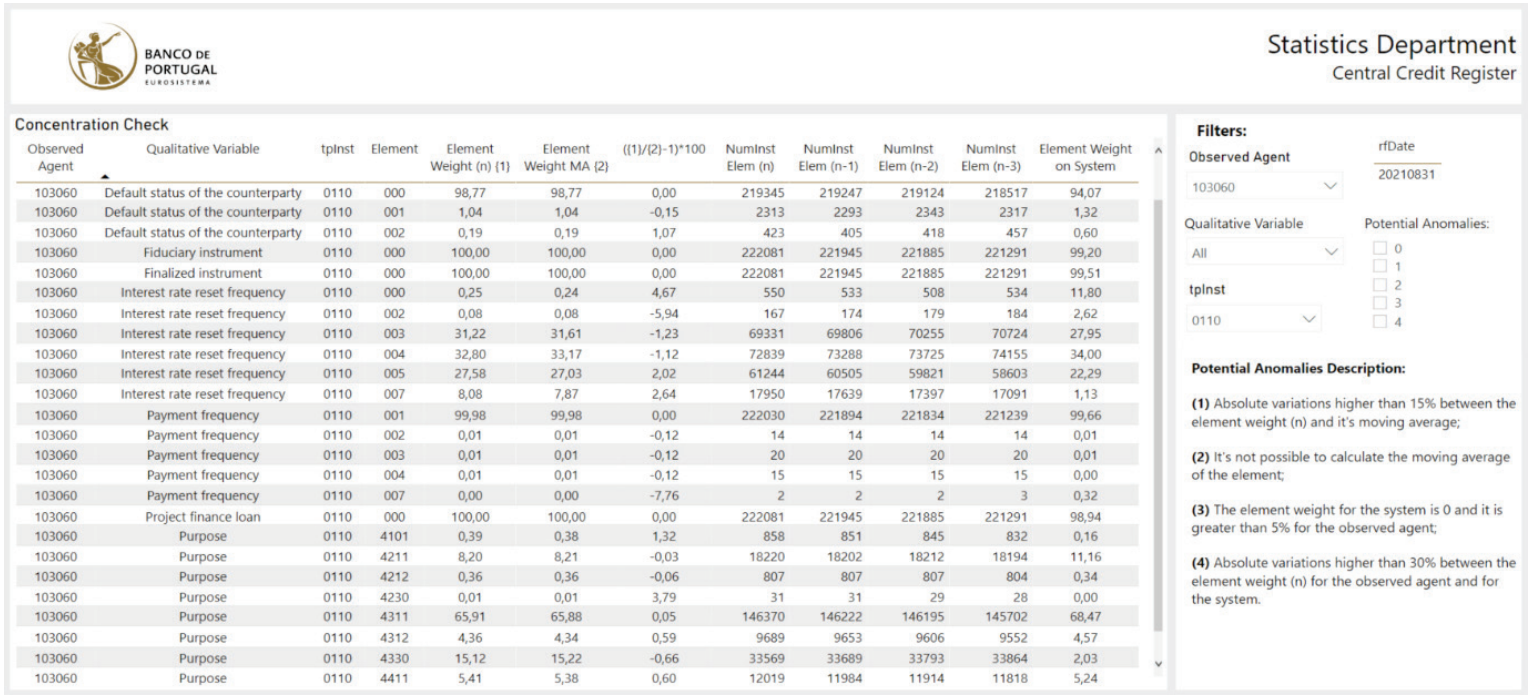
Reporting Consistency Test

- Only instruments with potentially anomalous patterns are displayed;
- The amounts associated with each instrument|debtor are presented to allow for measuring the overall impact of potential anomalies;
- Automatic filters were defined to allow the detection of potential anomalies such as:
  - Reporting gaps;
  - Lack of / inconsistent finalization.



Concentration Check

- The goal of this test is to evaluate the reporting of categorical variables;
- For each element we display:
  - The weight in the most recent period;
  - The weighted average of the previous three months;
  - The number of instruments for the last four months;
  - The average for the entire system in the most recent period.



BANCO DE PORTUGAL
EUROSISTEMA

Statistics Department
Central Credit Register

Concentration Check

Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	((1)/(2)-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Default status of the counterparty	0110	000	98,77	98,77	0,00	219345	219247	219124	218517	94,07
103060	Default status of the counterparty	0110	001	1,04	1,04	-0,15	2313	2293	2343	2317	1,32
103060	Default status of the counterparty	0110	002	0,19	0,19	1,07	423	405	418	457	0,60
103060	Fiduciary instrument	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	99,20
103060	Finalized instrument	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	99,51
103060	Interest rate reset frequency	0110	000	0,25	0,24	4,67	550	533	508	534	11,80
103060	Interest rate reset frequency	0110	002	0,08	0,08	-5,94	167	174	179	184	2,62
103060	Interest rate reset frequency	0110	003	31,22	31,61	-1,23	69331	69806	70255	70724	27,95
103060	Interest rate reset frequency	0110	004	32,80	33,17	-1,12	72839	73288	73725	74155	34,00
103060	Interest rate reset frequency	0110	005	27,58	27,03	2,02	61244	60505	59821	58603	22,29
103060	Interest rate reset frequency	0110	007	8,08	7,87	2,64	17950	17639	17397	17091	1,13
103060	Payment frequency	0110	001	99,98	99,98	0,00	222030	221894	221834	221239	99,66
103060	Payment frequency	0110	002	0,01	0,01	-0,12	14	14	14	14	0,01
103060	Payment frequency	0110	003	0,01	0,01	-0,12	20	20	20	20	0,01
103060	Payment frequency	0110	004	0,01	0,01	-0,12	15	15	15	15	0,00
103060	Payment frequency	0110	007	0,00	0,00	-7,76	2	2	2	3	0,32
103060	Project finance loan	0110	000	100,00	100,00	0,00	222081	221945	221885	221291	98,94
103060	Purpose	0110	4101	0,39	0,38	1,32	858	851	845	832	0,16
103060	Purpose	0110	4211	8,20	8,21	-0,03	18220	18202	18212	18194	11,16
103060	Purpose	0110	4212	0,36	0,36	-0,06	807	807	807	804	0,34
103060	Purpose	0110	4230	0,01	0,01	3,79	31	31	29	28	0,00
103060	Purpose	0110	4311	65,91	65,88	0,05	146370	146222	146195	145702	68,47
103060	Purpose	0110	4312	4,36	4,34	0,59	9689	9653	9606	9552	4,57
103060	Purpose	0110	4330	15,12	15,22	-0,66	33569	33689	33793	33864	2,03
103060	Purpose	0110	4411	5,41	5,38	0,60	12019	11984	11914	11818	5,24

Filters:

Observed Agent: 103060

rfDate: 20210831

Qualitative Variable: All

Potential Anomalies: 0, 1, 2, 3, 4

tpInst: 0110

Potential Anomalies Description:

(1) Absolute variations higher than 15% between the element weight (n) and it's moving average;

(2) It's not possible to calculate the moving average of the element;

(3) The element weight for the system is 0 and it is greater than 5% for the observed agent;

(4) Absolute variations higher than 30% between the element weight (n) for the observed agent and for the system.

- The purpose of the test is two-fold:
  - Checking the reporting consistency at the observed agent level;
  - Checking if there are significant differences in the reporting between the observed agent and the system as a whole.

Concentration Check - an example

- Up to July 2020:
  - Over 80% in the element “Other purposes” (6000) for housing credit (0110), usually a residual element.
- In August 2020:
  - The majority of the instruments were reallocated across the remaining elements, in particular to “Residential real estate purchase – permanent residential property” (4311).
  - The new weights for the observed agent are now in line with the system.
- In August 2021:
  - After this correction the reporting has been quite stable.

August 2020

Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Purpose	0110	4101	0,34	0,10	227,31	755	236	228	219	0,06
103060	Purpose	0110	4211	8,06	1,07	651,88	18062	2440	2379	2325	6,16
103060	Purpose	0110	4212	0,36	0,05	612,31	815	116	114	110	0,21
103060	Purpose	0110	4230	0,01	0,00	285,44	27	7	7	7	0,00
103060	Purpose	0110	4311	65,66	15,28	329,65	147055	34203	34209	34182	72,77
103060	Purpose	0110	4312	4,11	0,58	610,30	9211	1312	1295	1249	2,61
103060	Purpose	0110	4330	15,68	0,25	6.060,53	35120	573	566	567	2,01
103060	Purpose	0110	4411	5,08	0,02	21.273,61	11372	52	54	55	4,01
103060	Purpose	0110	6000	0,69	82,63	-99,17	1543	184936	184918	184902	0,42

August 2021

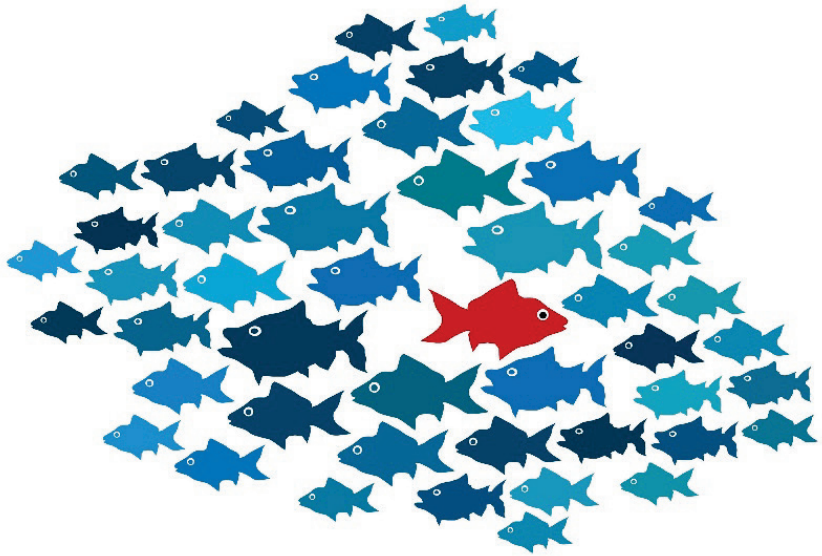
Observed Agent	Qualitative Variable	tpInst	Element	Element Weight (n) {1}	Element Weight MA {2}	({1}/{2}-1)*100	NumInst Elem (n)	NumInst Elem (n-1)	NumInst Elem (n-2)	NumInst Elem (n-3)	Element Weight on System
103060	Purpose	0110	4101	0,39	0,38	1,32	858	851	845	832	0,16
103060	Purpose	0110	4211	8,20	8,21	-0,03	18220	18202	18212	18194	11,16
103060	Purpose	0110	4212	0,36	0,36	-0,06	807	807	807	804	0,34
103060	Purpose	0110	4230	0,01	0,01	3,79	31	31	29	28	0,00
103060	Purpose	0110	4311	65,91	65,88	0,05	146370	146222	146195	145702	68,47
103060	Purpose	0110	4312	4,36	4,34	0,59	9689	9653	9606	9552	4,57
103060	Purpose	0110	4330	15,12	15,22	-0,66	33569	33689	33793	33864	2,03
103060	Purpose	0110	4411	5,41	5,38	0,60	12019	11984	11914	11818	5,24
103060	Purpose	0110	6000	0,23	0,22	4,07	518	506	484	497	5,96

Isolation Forest

- The Isolation forest is an unsupervised learning algorithm that works on the principle of isolating anomalies;

- Main advantages:

  - It has quasilinear complexity which makes it viable to integrate it in our daily routines;
  - It is a scoring algorithm allowing us to filter the most probable outliers when we are analyzing the results.



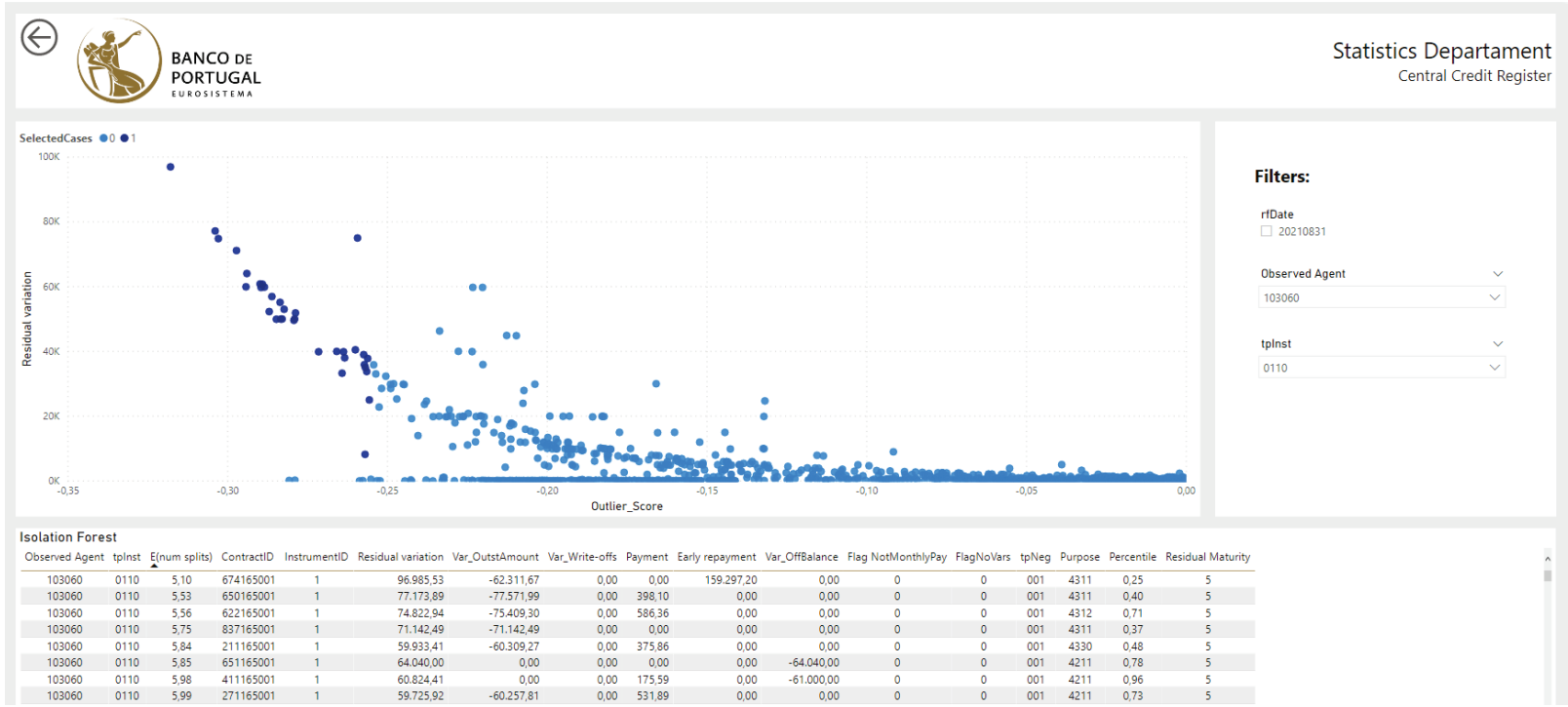
- The observations that require the least number of steps to be isolated will have a greater probability of being anomalies;

- The isolation forest is built using information from the four months prior to the most recent period to generate the training set.

Isolation Forest

- The goal is to detect abnormal evolutions of the outstanding amounts and associated qualitative variables;

- Model specification:
  - Residual variation[1];
  - Purpose of the loan;
  - Type of negotiation;
  - Residual maturity.



Observed Agent	tpInst	E(num splits)	ContractID	InstrumentID	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance	Flag NotMonthlyPay	FlagNoVars	tpNeg	Purpose	Percentile	Residual Maturity
103060	0110	5,10	674165001	1	96.985,53	-62.311,67	0,00	0,00	159.297,20	0,00	0	0	001	4311	0,25	5
103060	0110	5,53	650165001	1	77.173,89	-77.571,99	0,00	398,10	0,00	0,00	0	0	001	4311	0,40	5
103060	0110	5,56	622165001	1	74.822,94	-75.409,30	0,00	586,36	0,00	0,00	0	0	001	4312	0,71	5
103060	0110	5,75	837165001	1	71.142,49	-71.142,49	0,00	0,00	0,00	0,00	0	0	001	4311	0,37	5
103060	0110	5,84	211165001	1	59.933,41	-60.309,27	0,00	375,86	0,00	0,00	0	0	001	4330	0,48	5
103060	0110	5,85	651165001	1	64.040,00	0,00	0,00	0,00	0,00	-64.040,00	0	0	001	4211	0,78	5
103060	0110	5,98	411165001	1	60.824,41	0,00	0,00	175,59	0,00	-61.000,00	0	0	001	4211	0,96	5
103060	0110	5,99	271165001	1	59.725,92	-60.257,81	0,00	531,89	0,00	0,00	0	0	001	4211	0,73	5

- The criteria to select the observations in dark blue consisted in three cumulative conditions:

  Be from the 1% more negative outlier scores | Have an outlier score of -0.25 or less | Have a residual of 1,000€ or higher.

[1] Residual Variation = $Abs(\Delta\text{Outstanding nominal amount} + \Delta\text{Accumulated write-offs} + \text{Payment} + \Delta\text{Off-balance-sheet amount} + \text{Early repayment})$

Isolation Forest

- Multiple situations can generate significant outlier scores, either due to the residual variation:
  - Monthly payment without a decrease of the outstanding amounts (1) or they are significantly different (3);
  - Early repayment without a similar decrease in the outstanding amounts (2);
  - The off-balance sheet amount decreases without an increase of the outstanding amounts (4);
  - Decrease in write-offs, with no increase in the outstanding amounts (5).

Observed Agent	tpInst	E(num splits)	ContractID	InstrumentID	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance
105070	0110	5,05	045823733	1	130.100,00	0,00	0,00	130.100,00 (1)	0,00	0,00
103060	0110	5,10	674165001	1	96.985,53	-62.311,67	0,00	0,00	159.297,20 (2)	0,00
103060	0110	5,53	650165001	1	77.173,89	-77.571,99	0,00	398,10 (3)	0,00	0,00
103060	0110	5,85	651165001	1	64.040,00	0,00	0,00	0,00	0,00	-64.040,00 (4)
103060	0110	6,01	130165001	1	60.738,15	0,00	-60.738,15 (5)	0,00	0,00	0,00

- Or due to the uncommon combination of the three qualitative variables

Outlier_Score	Residual variation	Var_OutstAmount	Var_Write-offs	Payment	Early repayment	Var_OffBalance	tpNeg	Purpose	Percentile	Residual Maturity
-0,2570	8.202,70	-8.202,70	0,00	0,00	0,00	0,00	001	4311	0,02	-1: Passed
-0,2557	25.005,00	0,00	0,00	0,00	0,00	-25.005,00	001	4101	0,95	5: Over 20 years

Main conclusions

- The new tests have been used in production, on a monthly basis, for over a year with encouraging results;
- They have allowed us to:
  - Detect reporting gaps and strange patterns, even subtle ones that only affect a few instruments;
  - Oversee the evolution of the reporting of qualitative variables and to detect structural changes;
  - Monitor the evolution of the amounts reported instrument by instrument, taking into consideration the qualitative variables that characterize them and ranking them by the degree of severity for further questioning.
- Hence, these new tools have contributed to an increase both in the effectiveness and efficiency of the data quality assessment process and so far we have received very positive feedback from the analysts.