IFC International Finance Corporation WORLD BANK GROUP

Case Study

Leveraging Machine Learning *to* Enhance Credit Data Quality: Case Study *of* AI use *by the* Central Bank *of* Brazil's Public Credit Registry[1]

MAY 2025

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

ICCR

WORLD BANK GROUP
THE WORLD BANK IBRD • IDA
IFC International Finance Corporation

1 Introduction

The proliferation of technological advancements and innovations over the past decade are reshaping the credit reporting industry. Advanced computing, artificial intelligence, big data (alternative data) and biometric identification, based on sound regulatory and legal foundations, can facilitate improved financial inclusion by providing lenders with the information and systems they need to confidently lend. In turn, these systems can provide businesses and borrowers with better credit terms and greater transparency into the processes and metrics by which they can access financing. Increased innovation allows credit information systems to access, collect, aggregate, manipulate, and analyze large quantities of data from new sources in a faster and more efficient manner. Such innovations have enabled greatly improved identification, analytical, and other capabilities.

In response to these changes, the International Committee on Credit Reporting (ICCR), released a white paper in 2022 on the responsible use of technology in credit reporting. The paper presents 10 principles on how technology can be used responsibly, in a manner that minimizes any unintended negative outcomes of these innovations.

While private sector entities, such as credit bureaus, have been at the forefront of adopting the new technologies, there is a belief that public sector entities are slow to adopt technological innovations. This perception is mainly driven by the needs of public sector entities to balance innovation with potential impacts on the public sector, and because of perceived costs in implementation. However, evidence has shown that public sector authorities are exploring the introduction of new technology where it makes sense, and where it can empower processes like quality assurance and supervision.

An estimated 50% of businesses worldwide, including small and large firms, have adopted specifically machine learning (ML) to enhance their operations. Adoption rates are high for developed economies and finance industry. For example 72% of the US enterprises reporting using ML, while the finance industry are using ML for fraud detection. Adoption of ML lags for public sector, even for developed markets.

The objective of the paper is to demonstrate how public sector entities are embracing ML learning in their processes. The case study explores the use of artificial intelligence (AI), specifically machine learning (ML), by Brazil's public credit registry, known as the Credit Information System (SCR[2]), of Brazil's Central Bank (BCB[3]) to enhance data quality assurance and data governance. The paper describes the background of the SCR, the use case and methodology, and findings and future opportunities. The paper concludes by sharing key lessons for other public sector authorities considering using AI.

1 Prepared by Collen Masunda (IFC), Rogerio Rabelo Peixoto and David Paulo Pereira (Banco Central do Brasil).

2 Sistema de Informações de Crédito.

3 Banco Central do Brasil.

2

Overview of the Credit Information System (SCR)

The SCR is a public credit registry system developed and maintained by Banco Central do Brasil (BCB) to register credit operations, providing detailed information about borrowers and loans. The SCR supports financial institutions in assessing the creditworthiness of individuals and companies, enabling them to make informed decisions on loan approvals, credit limits, and interest rates.

As a traditional public credit register, the SCR plays a vital role in BCB's responsibility for overseeing prudential credit risk. It allows BCB to monitor the aggregated level of indebtedness within the economy, detect potential risks in credit exposure, and take timely actions to prevent systemic threats to the financial system. The SCR consolidates detailed credit information about all credit operations, from banks, credit unions, and other lenders in Brazil. It processes 1.1 billion credit operations monthly, which include detailed information about loan agreements, outstanding debts, and lines of credit across the country. The SCR identifies a total of 141 million distinct individuals with outstanding debt above 200 Brazilian reais (R$) (~US$33), as well as 9 million distinct companies. Data are submitted regularly by 1,233 financial institutions and 2,264 fixed-income investment funds backed by credit receivables. The SCR's analytical database (data warehouse) maintains historical records dating back to 2003, representing over 22 years of credit activity. These operations consume more than 100 TB (terabytes) of disk space, highlighting the system's big scale and central role in Brazil's financial infrastructure.

Ensuring data accuracy and completeness within SCR is thus essential for effective credit granting and regulatory oversight and decision-making.

3 Traditional Ways of Identifying Anomalies

BCB faces several challenges when it comes to assessing and ensuring the quality of data received from financial institutions. Given the large volume of transactions processed monthly, traditional statistical methods for anomaly detection face limitations. Some of the challenges are highlighted below:

- Data Volume: The large volume of credit data presents a challenge for BCB in terms of processing, storage, and analysis. This amount of information requires advanced systems and infrastructure to ensure timely and accurate data quality management.
- Data Accuracy and Completeness: The SCR data quality depends heavily on the accuracy and completeness of the information provided by financial institutions. BCB relies on these institutions to report information correctly and in a timely manner.
- Fraud Detection: BCB must also deal with the risk of fraud and manipulation of credit data. Financial institutions may sometimes report inaccurate information, either intentionally or unintentionally, due to operational errors or fraudulent activity. Identifying and preventing these issues is critical to ensure the credibility of the SCR.
- Data Timeliness: The SCR relies on monthly data for decision-making. Delays in data reporting can create gaps in the credit information, making it difficult for BCB to monitor credit risk or detect emerging trends in the economy. The challenge lies in maintaining an efficient, timely process that allows accurate and up-to-date data to be consistently available.

Given these potential issues, it is crucial for BCB to implement enhanced data validation systems and processes to identify anomalies in loan data as early as possible. Traditional statistical methods, such as Z-Score (Altman 1968) and Interquartile Range (IQR) (Tukey 1977), have long been used for anomaly detection, but they exhibit limitations when applied to large, high-dimensional, and evolving datasets like those found in the SCR.

4 Leveraging Machine Learning (ML) to Improve Data Quality

To address the limitations of traditional methods, BCB decided to explore the efficacy of machine learning models for detecting anomalies and ensuring data accuracy and completeness. The SCR contains multiple interdependent variables, such as loan terms, interest rates, and disbursed amounts. Some of the key considerations that influenced the use of ML are shown in Table 1.

Table 1

Key Considerations Influencing the Use of ML by BCB.

	Traditional statistical methods.....	Machine learning models.....
Adaptability to high-dimensional data	struggle to account for relationships between multiple interdependent variables.	can analyze multidimensional interactions and identify deviations more effectively.
Robustness to data distribution variations	require well-defined distributions to set threshold limits, making them sensitive to skewed data.	dynamically adjust to changing distributions, enhancing their ability to identify true anomalies.
Scalability to large datasets	are problematic because manual rule-based anomaly detection is difficult in large data sets.	can continuously learn from historical data, automate the detection process, and scale seamlessly to accommodate increasing data loads without compromising performance.

BCB decided to pilot an unsupervised machine learning approach that identifies anomalies in key loan data. The pilot was designed to improve data accuracy for four data fields that define fundamental loan agreements:

- Loan interest rate
- Disbursed amount to the borrower
- Debtor's monthly income
- Loan term (in days)

Ensuring that these critical fields are accurate is essential for maintaining the overall integrity of the SCR database. Incorrect values in fields can arise from various sources, including manual entry errors, system integration issues, or data migration problems. For example, unusually high or low interest rates for a specific credit product could indicate potential data quality issues. Similarly, disbursed amounts that are inconsistent with the debtor´s income could suggest incorrect reporting, potentially pointing to fraud or other operational problems. Another example of a data quality issue arises when loan terms are reported incorrectly—whether through longer or shorter repayment periods that deviate significantly from the norm.

By adopting modern ML approaches, BCB expects to significantly improve the accuracy and completeness of the SCR database, ensuring that it remains a reliable tool for assessing credit risk and managing systemic threats. It is anticipated that the transition from traditional statistical methods to advanced AI-driven validation will enhance the robustness of credit data and strengthen the integrity of Brazil's financial system.

Methodology

To address data quality assurance weaknesses caused by increasingly large volumes of data, BCB implemented the Isolation Forest (iForest) algorithm (Liu et al. 2008), an unsupervised ML approach designed for anomaly detection (also referred to as outlier detection). The iForest model offers several advantages over other outlier detection methods, primarily due to its efficiency and scalability. It is notably fast, operating in linear time, which makes it suitable for processing large datasets quickly.

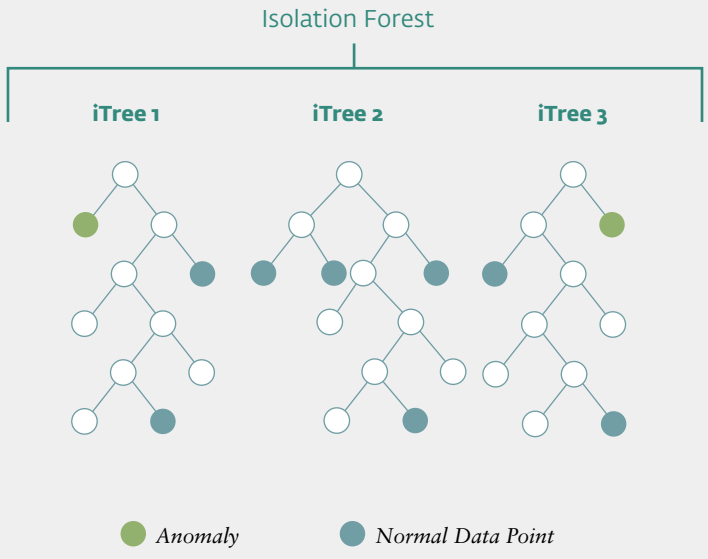
Additionally, its scalability is supported by a small memory footprint, enabling it to manage high-volume data environments effectively. Furthermore, the algorithm is simple and easy to understand, facilitating interpretation of results. Unlike other statistical outlier detection methods, iForest requires no assumptions about the underlying data distribution, allowing it to be applied broadly across diverse datasets without the need for prior adjustments or transformations.

iForest works by constructing an ensemble of random decision trees, or iTrees, which are used to partition the dataset. The key intuition is that outliers are few and different, and thus should be easier to isolate compared to normal data points. Each tree in the forest is built by randomly selecting a feature and then randomly choosing a split value between the minimum and maximum value of the selected feature. This process of random partitioning continues recursively, creating branches until the data points are isolated. Outliers, being distinct and sparse, tend to be isolated quickly, requiring fewer splits on average compared to normal points (Fig. 1).

Fig. 1

Isolation Forest iTrees

Source: SpotIntelligence.com



The path length from the root to the terminating node for each data point in all trees is averaged, and points with shorter average path lengths are considered outliers. This averaging across an ensemble of trees provides a robust outlier score.

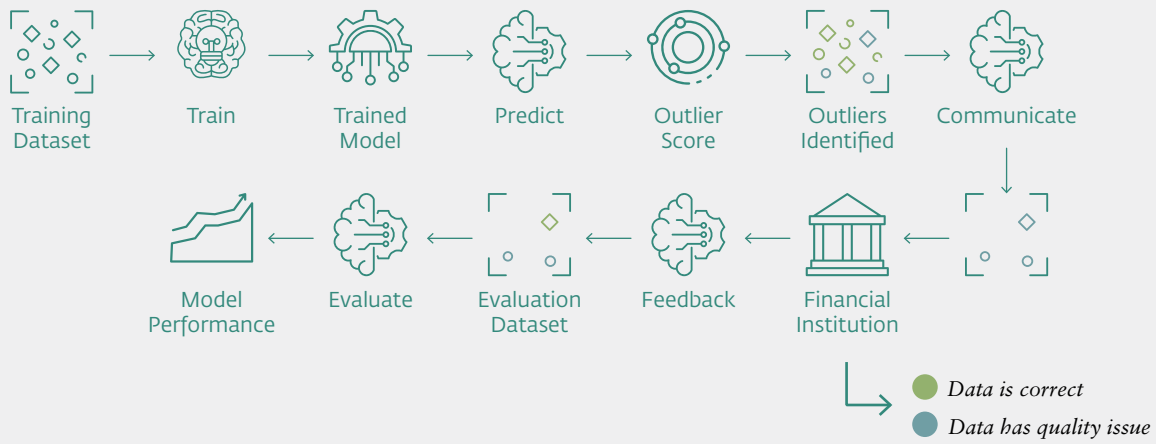
BCB implemented an outlier detection cycle, which leverages the iForest model, to systematically identify outliers in paycheck-linked loans. The procedure (Fig. 2) consists of the following steps:

- Training: First, BCB trained a model using a paycheck-linked loan training dataset, which consists of over 55 million credit operations. Using the Python programming language and the Scikit-Learn machine learning library, BCB built the model focusing on four data fields that encapsulate the essential loan properties: interest rate, loan term, disbursed amount, and the debtor's income. The model's hyperparameters were configured to generate 100 trees, with each tree built by sampling 10% of the training dataset, thus ensuring diversity and robustness in detecting different outlier patterns.
- Prediction: Once trained, the model was deployed across the entire paycheck-linked loan dataset to compute an outlier score for each loan. These scores represent the likelihood of an observation being anomalous. By using a decision threshold score, each loan was classified as either an outlier or a normal observation.
- Communication: Subsequently, the SCR quality warning system flags the outliers to the financial institutions. Institutions are required to investigate their list of flagged loans, examining whether the outliers found are indeed quality issues, or whether they were falsely identified by the model.
- Feedback: Financial institutions must inform BCB of their findings for each flagged loan. They report whether a correction is warranted for any of the four fields reported, or if the loan's data is accurate and the outlier status is indeed a false positive. This feedback is crucial, as it provides the ground-truth labels essential for iterative model refinement.
- Evaluation: The insights and ground-truth labels garnered from the feedback step play an important role in improving the model. By interactively building an evaluation dataset after each cycle, BCB can continuously reevaluate the model's performance.

Fig. 2

BCB Process Overview

Source: Authors



Findings

Our outlier identification model, specifically tailored for paycheck-linked loans, was evaluated using the precision metric. Precision is a performance metric used in classification tasks to measure the accuracy of positive predictions. It is calculated as:

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)}$$

A high precision score indicates fewer false positives and minimizing the false positive rate is critical. Communicating outliers requires financial institutions to analyze data and respond promptly, and excessive false positives not only impose unnecessary costs but also undermine the credibility of the process.

The initial model identified 278 paycheck-linked loans as potential outliers across nine financial institutions. BCB received feedback on 206 of these cases from eight institutions, confirming that 179 loans exhibited genuine data quality issues, while 27 were false positives. This resulted in a precision rate of 87%. A detailed analysis of the false positives revealed that most involved high-income debtors earning over R$30,000 (~US$5,260) per month or loan disbursements exceeding R$500,000 (~US$87,600).

Based on this insight, BCB refined its approach by developing two distinct models: one tailored for high-income debtors and another for the remaining debtors. After redesigning, retraining, and reevaluating the model using the validated dataset, the precision rate improved significantly to 97%.

The iForest model effectively identified a range of quality issues (Table 2), including anomalous interest rates significantly deviating from the dataset norm and disbursements inconsistent with debtors' income.

Table 2

Examples of outliers identified by iForest

Interest Rate (%)	Term (days)	Disbursed Amount (R$)	Income (R$)
49.36	401	4,200.00	2,002,091,101,342.95
25.64	1,484	7,500.00	2,002,091,105,455.94
9,999.00	57	535.21	3,545.94
4,315.42	111	810.73	1,595.00
6,364.39	204	341.29	1,326.25
0	25,159 (68 years)	45,485.00	4,000.00
12.68	17,427 (47 Years)	7,358.87	2,000.00
11.33	3,679	957,425.00	6,300.00
7.53	7,295	548,425.80	10,000.00
0	1,323	1,237,137.80	30,700.00

5 Current Status and Future Work

The noteworthy results achieved using the iForest model highlight the potential for further exploration and diversification of ML tools to strengthen the SCR quality process. Future iterations of this project may involve integrating other outlier detection models, such as One-Class Support Vector Machines (SVM) (Schölkopf et al. 2001), Local Outlier Factor (Breunig et al. 2000), and K-Means clustering (MacQueen 1967).

The table below presents a comparison of the key characteristics of the three models:

Model	Detection Type	Supervision	Strengths	Use in Quality Process	Use in Fraud Detection
One-Class SVM	Anomaly Detection	Unsupervised (trained on normal data)	Effective in high-dimensional spaces; good for rare anomaly detection	Detects deviations in loan features and reporting standards	Flags synthetic identities, ghost loans, and misreporting
Local Outlier Factor (LOF)	Anomaly Detection	Unsupervised	Detects local anomalies based on density; good for contextual outliers	Identifies local inconsistencies in loan data and reporting	Detects local fraud patterns like loan stacking or ghost loans
K-Means Clustering	Clustering & Outlier Detection	Unsupervised	Efficient clustering; useful for segmentation and detecting cluster outliers	Segments loan portfolios and benchmarks institutions	Identifies suspicious borrower clusters and structuring patterns

Additionally, advanced deep learning techniques, such as Recurrent Neural Networks (RNNs) (Rumelhart et al. 1986) and particularly Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber 1997), could be explored to improve pattern recognition in temporal credit data. These models can capture long-term dependencies, making them valuable for identifying temporal inconsistencies, especially over outstanding debt amounts.

Additionally, advanced deep learning techniques, such as Recurrent Neural Networks (RNNs) (Rumelhart et al. 1986) and particularly Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber 1997), could be explored to improve pattern recognition in temporal credit data. These models can capture long-term dependencies, making them valuable for identifying temporal inconsistencies, especially over outstanding debt amounts.

To further enhance the process, some key challenges must be addressed:

- Automating the Feedback Step: Currently, feedback is provided in unstructured natural language, making the analysis challenging for the SCR quality team. Implementing a structured feedback CSV document would streamline data processing and improve scalability.
- Enriching Feedback Information: Financial institutions should provide more detailed explanations when reporting false positives, highlighting specific business factors that made these loans stand out from others in the market. This additional context would help refine and improve our outlier detection models.
- Strengthening Process Governance: Encouraging greater cooperation and ensuring timely feedback from financial institutions will enhance the reliability of our quality process.

Furthermore, extending the application of these models beyond paycheck-linked loans to other credit products such as credit cards, real estate financing, and car loans would enable a more comprehensive approach to outlier detection. Each product has unique characteristics, presenting an opportunity to harness ML techniques for broader quality assessment.

6

Lessons for Other Public Sector Entities Considering Testing New Technologies

Traditionally, regulatory entities are viewed as risk averse and take a long time to adopt technology. Key factors to consider for public sector entities seeking to undertake a similar exercise:

a. Support from the top—BCB is an innovative central bank and has pioneered several innovations to modernize the financial system, promote competition, and foster financial inclusion. Key initiatives include Pix (BCB a), a real-time payments system with nationwide adoption; Drex (BCB b), the upcoming central bank digital currency aimed at enabling tokenized assets and smart contracts; and Open Finance (BCB c), which allows consumers to share financial data across institutions to access more tailored services.

b. Defining clear use case: the objective was clearly spelled out and was at the core of the work of the SCR. It is important to focus on use cases answering problems that several stakeholders can relate with.

c. Stakeholder buy-in given the nature of the use case, the BCB had to secure both internal and external buy-ins. The latter was important to validate the precision rate and possibility for the evolution of the model training.

d. Incremental testing: BCB decided to start with just four variables included in the SCR and to build on the learnings as they scaled.

e. Robust information technology (IT) infrastructure: To support the development and operation of AI solutions, BCB has launched two key infrastructure initiatives. The Laboratory of Analytical Intelligence (LIA4) is a collaborative platform designed for data analysis applications, particularly in Python. It enables users to perform large-scale exploratory data analysis and ML experiments using either a VS Code interface (Microsoft) or JupyterLab notebooks (Jupyter) running on high-performance virtual machines. Complementing LIA, the internal BCB cloud infrastructure provides a modern cloud environment for deploying enterprise-grade applications. By leveraging Docker containerization (Docker) and Kubernetes orchestration (Kubernetes), BcCloud ensures high availability, scalability, and performance for AI and ML solutions within BCB.

f. Robust process governance: The Center of Excellence in Data Science and Artificial Intelligence (CdE IA5) is a newly established community of technical experts who meet regularly to share best practices and use cases related to artificial intelligence within the BCB. CdE IA is expected to play an important role in advancing AI adoption by proposing governance guidelines to ensure the safe and ethical development of AI applications. Additionally, it will support workforce development by updating training programs and recommending requirements for the responsible implementation of generative AI tools and services.

g. Transparency: it is important to be transparent about the objectives, methodology, and expected implications of any proposed technological innovations.

4 Laboratório de Inteligência Analítica

5 Centro de Excelência de Ciência de Dados e Inteligência Artificial

7 Conclusions

Central banks and public credit registers should consider opportunities for implementing smart innovations, not only to keep pace with their private partners, but also to enhance their processes and service delivery.

The adoption of ML for data quality assurance by BCB demonstrates significant opportunities to enhance the accuracy of the SCR data it receives from financial institutions. By leveraging outlier detection models like iForest, BCB can automate the identification of anomalies, leading to more efficient and scalable detection of data quality issues.

Moreover, this case study reinforces the strategic value of artificial intelligence and machine learning in strengthening public sector capabilities. By adopting these technologies, public institutions can not only improve operational efficiency and data governance but also foster a culture of innovation that aligns with the pace of digital transformation seen in the private sector.

It is important that the design of the innovations should meet the basic tenets of responsible use of technology such as ensuring model interpretability and sustaining model performance in the face of evolving data, highlighting the need for ongoing research and adaptation. Continuous refinement of models is essential to maintain their relevance and effectiveness.

Encouraging the responsible and transparent use of AI in public institutions can lead to more robust supervisory frameworks, better-informed policy decisions, and ultimately, greater trust in public financial infrastructure. As the financial ecosystem becomes increasingly data-driven, it is imperative that public authorities embrace these tools—not as optional enhancements, but as essential components of modern governance and oversight.

This experience from Central Bank of Brazil serves as a compelling example for other public sector entities worldwide, that with the right safeguards and strategic vision, AI can be a powerful ally in delivering more accurate, inclusive, and resilient financial systems.

References

Altman, E. I. 1968. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." The Journal of Finance, 23(4): 589–609.

BCB. a. Pix – Instant payments. Retrieved April 15, 2025, from https://www.bcb.gov.br/en/financialstability/pix_en

BCB. b. Drex – Digital Brazilian Real. Retrieved April 15, 2025, from https://www.bcb.gov.br/en/financialstability/drex_en

BCB. c. Open Finance. Retrieved April 15, 2025, from https://www.bcb.gov.br/en/financialstability/open_finance

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. 2000. "LOF: Identifying density-based local outliers." Proceedings of the ACM SIGMOD International Conference on Management of Data: 93–104.Docker. Retrieved April 15, 2025, from https://www.docker.com

Hochreiter, S., & Schmidhuber, J. 1997. "Long short-term memory." Neural Computation, 9(8): 1735–1780.

Jupyter. Retrieved April 15, 2025, from https://jupyter.org/

Kubernetes. Retrieved April 15, 2025, from https://kubernetes.io

Liu, F. T., Ting, K. M., & Zhou, Z. H. 2008. "Isolation Forest." Proceedings of the 2008 IEEE International Conference on Data Mining (ICDM): 413–422. https://ieeexplore.ieee.org/document/4781136

MacQueen, J. 1967. "Some methods for classification and analysis of multivariate observations." Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1: 281–297.

Microsoft. Retrieved April 15, 2025, from https://code.visualstudio.com

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986. "Learning representations by back-propagating errors." Nature, 323(6088): 533–536.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. 2001. "Estimating the support of a high-dimensional distribution." Neural Computation, 13(7): 1443–1471.

Tukey, J. W. 1977. Exploratory Data Analysis. Addison-Wesley

About IFC

IFC—a member of the World Bank Group—is the largest global development institution focused on the private sector in emerging markets. We work in more than 100 countries, using our capital, expertise, and influence to create markets and opportunities in developing countries. In fiscal year 2024, IFC committed a record $56 billion to private companies and financial institutions in developing countries, leveraging private sector solutions and mobilizing private capital to create a world free of poverty on a livable planet. For more information, visit www.ifc.org.