NTT DATA Services

WHITE PAPER | BUSINESS PROCESS OUTSOURCING

Data Anomalies in the Digital World Are Real — Can Your Bank Detect Them Successfully?

MARCH 2020

NTT DATA

Introduction

Using anomaly detection to root out fraud has been around for a long time. Myriad options are available for financial institutions, but they can be costly and complex to implement. This paper outlines the benefits and limitations of each method and examines how a hybrid approach can be used to make data anomaly detection manageable.

Financial institutions leverage anomaly detection for a wide range of uses, including:

- Discovering unusual transactions or activities in accounts, such as:
  - Sudden surges in transactions in an otherwise dormant account
  - Higher value transactions than usually occur in an account
  - Transactions in another country — for example, payments made from a U.S. account within another country
  - Other fraudulent activities with a deliberate and malicious intent
- Detecting synthetic IDs created via bot factories
- Monitoring watch lists against accounts and transactions, such as sanctions and high-risk customers in know-your-customer (KYC) processing stages
- Identifying the manipulation of claims processing in the insurance industry
- Detecting anomalies in invoice processing — for example, value-added tax mismatches or pricing and discount variations

Watch out for these common anomalies

False positives

In a false positive or "false alarm" scenario, an anomaly is detected based on preset rules but turns out to be a legitimate transaction. For example, a one-time payment of $1 million made to an individual's account could be the result of the sale of an asset from a property inheritance.

The impact of a false positive can be very damaging to a financial institution's business. It could lead to customer friction, increased operational costs and, potentially, reputational damage. Organizations need a way to ensure that false positives are filtered out from all identified anomaly data.

False negatives

A false negative is the opposite of a false positive. In this scenario, no alarms are detected or triggered, and the transaction completes successfully, passing through all prebuilt security and other defenses.

It's difficult to detect false negatives because they bypass the organization's defenses. A commonly experienced false negative occurs with manufactured synthetic IDs.

Point

A data instance that deviates from the dataset's normal pattern can be considered an anomaly. For example, if fuel for an automobile costs $51 per day but increases to $501 on any random day, then it's an anomaly.

Contextual

If a data instance behaves anomalously in one context but not in another context, then it's considered a contextual — or conditional — anomaly. For example, credit card charges are usually higher during Christmas than they are the rest of the year. And although high, these may not be anomalous because higher charges are contextually normal at Christmastime. Conversely, an equally high credit card bill at a random non-holiday time of year could be a contextual anomaly.

Collective

When a group of similar data instances behaves anomalously compared to the whole dataset, it's considered a collective anomaly. And while an individual data instance may not be an anomaly by itself, when it's part of a collection it could be identified that way. For example, an individual debit or credit card charge of less than $100 in a day may not be classified as an anomaly in a normal scenario, but if it's part of series of transactions totaling $1,000 on that particular day then that charge is a potential collective anomaly.

Evaluating commonly used anomaly detection approaches

It's important to have a basic overview of anomaly detection approaches, as each method has benefits and limitations. Depending on the scenario, many methods may apply — and in some cases more than one. Factors like implementation cost, availability of the right data and lack of domain knowledge to interpret the data can play an important role in selecting anomaly detection methods. Financial institutions have a lot of options for detecting anomalies.

Rule-based detection

Defining rules is probably the simplest and easiest way to implement anomaly detection. For example, applying a dollar value or other limit, like a single transaction greater than $1 million or a large number of transactions in one day greater than a certain value.

However, the limitations of a rule-based system become apparent when more complicated models are needed. For example, it's difficult and inefficient if we try to model abnormal transaction patterns relative to the rest of the transactions in a specific user segment. So, too, is tracking patterns such as sudden surges in transactions across a specific period of time — for example, the middle of the night or at month or quarter end.

Another issue with a rule-based approach arises when new anomalies or rules are identified. In these scenarios, the new anomaly or rule must be incorporated into the existing detection approach, necessitating constant monitoring and updates to the company's rules engine. In other words, this approach lacks the capability to "learn" from data sets and previous transactions. Rule-based systems also fail to identify behavior patterns across transactions and detect anomalies based on that information.

Nearest neighbor detection

This method, also known as distance-based detection, works on the underlying assumption that any new anomaly is close to other known anomalies. That is, the closer the new anomaly is to an existing anomaly, the higher the chances of it being classified as an anomaly. This nearest neighbor method requires organizations to first model a set of known anomalies; in other words, "supervised" learning needs to happen.

Although much better than a rule-based approach, the nearest neighbor method still has some of the same disadvantages — including the inability to "learn" by itself. So, any anomaly detection system leveraging this approach must also have behavior patterns fed into it if patterns are to be modeled and the system is to "learn."

Clustering detection

Clustering, or density based, is by far one of the most effective detection methods. These systems try to not just identify anomalies based on previously fed patterns but also learn new patterns based on the clusters. This makes the approach very effective, as it depicts and identifies anomalies in a real-time situation. Patterns can also be used to model behavior over a period of time, which both the rule-based and nearest neighbor models lack.

Isolation forest detection

Different from both the nearest neighbor and clustering approaches, the underlying principle in an isolation forest is the possibility of an individual instance (anomaly) being "isolated" from the rest of the instances. A detailed description is available from Fei Tony Liu, Kai Ming Ting and Zhi-Hua Zhou.[1]

Efficiency is a concern in any anomaly detection system. The following factors may affect it:

Domain-specific knowledge

Using domain-specific knowledge only to improve predictions is a mistake. If the goal is to gather possible causes, this will be a problem because predictions have nothing to do with causality. For example, predicting the percentage increase in the mortgage applications in a specific sector like agriculture may be linked to weather conditions, government policies for that year and/or the economic condition of the country, all of which could be common influencers.

Dynamic environments

Another area where anomaly detection is likely to fail is dynamically changing environments — where the definition of normal behavior changes. This is quite common in political and business environments, where change is so rapid that it's difficult to create a baseline against which anomalies can be measured. In any changing, dynamic environment, what isn't currently perceived as an anomaly may be classified as an anomaly in the future, and vice versa. For example, a sudden change in the political environment in a country, such as a change in leadership, could result in drastic regulatory policy changes in the financial industry that in turn could result in changing consumer behaviors — customers withdrawing deposits and resorting to other modes of investment due to the fear of uncertainty or housing loans declining at a rapid pace due to government policy changes. It's difficult to model such scenarios and identify real anomalies.

Fully automated anomaly detection approaches and solutions

Fully automated solutions are limited in anomaly detection. Many of the solutions available in the market are designed to reduce manual intervention completely and instead enable straight through processing. The main benefit to this type of anomaly detection system is cost. But factors beyond mathematical models need to be considered, too, many of which the fully automated solutions will be unable to capture. The various scenarios mentioned in the previous sections make a fully automated solution difficult to achieve in the real world.

An anomaly detection model's inability to incorporate human sentiments is another factor that makes many of the current automated solutions fail. While the models and data would accurately detect an anomaly, and that anomaly would be considered legitimate from all aspects — including legal — there could be situations where everything needs to be overridden and human judgement and compassion considered instead. For example, if a financial crisis occurs in a country and it leads to an anomalous situation like drastic regulatory policy changes impacting consumer behaviors, then the broader circumstances of that crisis must be considered and, in many cases, must override what a detection model provides as an insight.

Why a hybrid approach works best

Many approaches successfully reduce the number of false positives, significantly changing the way anomaly detection models are trained. Traditional methods base the model on a "bad" definition, meaning identified and/or known cases of fraud, non-compliance with the Bank Secrecy Act and other disruptive transactions.

Newer approaches train the model on a targeted definition of known "good" behaviors. The premise is that the known/identified "bad" population used for model training is based on a smaller sample size of the overall population — less than 5% (if not, then there is a larger issue with the models) — while the identified "good" population is greater than 95%. The "good" provides a more comprehensive population size on which to train the model, as well as a way to capture seasonality, maturity level of the portfolio, consumer behavior, and both macroeconomic impacts and the impact of previous management actions.

But trying to devise a solution that considers all factors may result in a system too complex and cost prohibitive to implement. It may also be very difficult to implement operationally. The best method is to adopt a hybrid approach that uses anomaly detection methods to get the initial subset of data, and then successfully filters out false positives to derive a subset of the initial list that can be used for further evaluation.

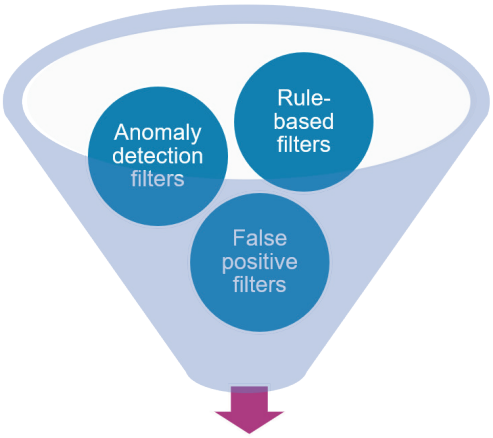


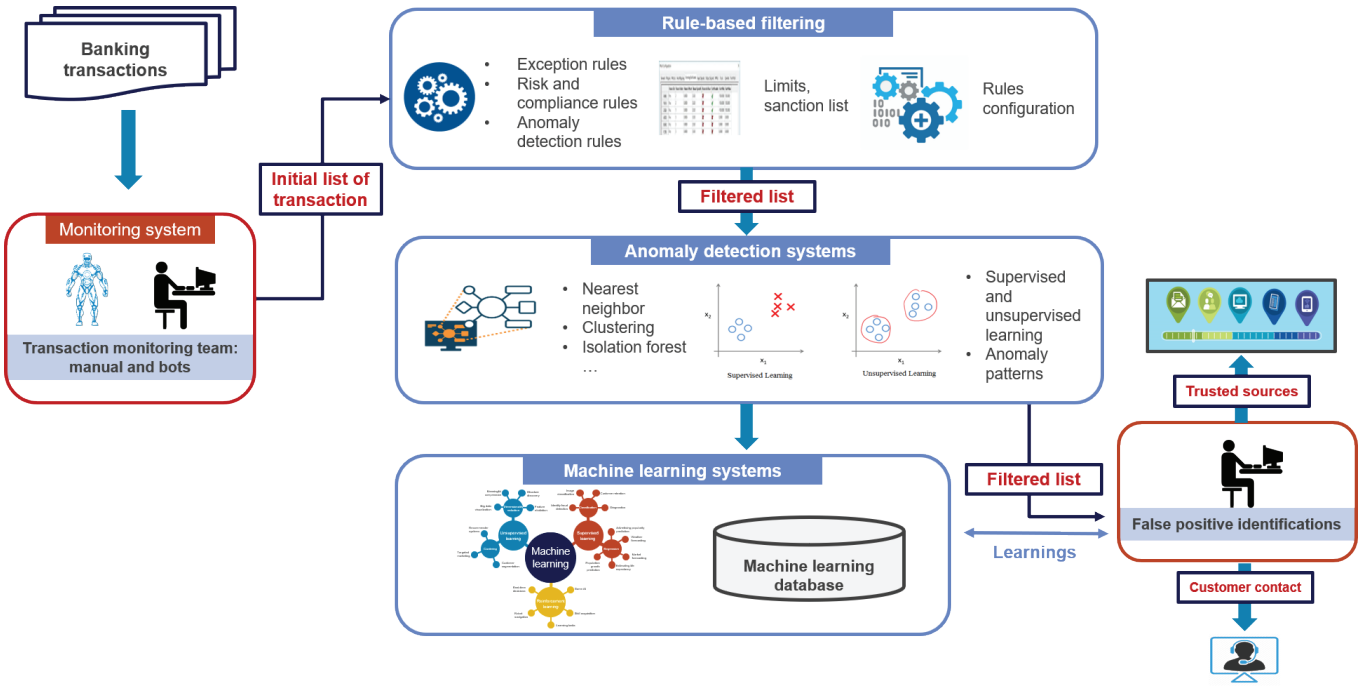
Figure 1: Hybrid approach to anomaly detection

The hybrid approach starts by running the entire dataset through rule-based filtering. The results are fed into an anomaly detection engine, which reduces the dataset into a smaller subset of transactions. That filtered list is fed into an anomaly detection filter — a platform based on an analytical model capable of identifying false positives and other anomalies using the various methods described in the sections above. This step reduces the dataset into an even smaller subset of transaction candidates for additional manual evaluation. A team processes this smaller subset manually, verifying false positives using methods such as document verification, as well as KYC information from existing customers. Customers may also be contacted directly for additional information and verification. This step results in a smaller number of transactions, some of which could be real anomalies that will need further investigation.

One additional filter worth mentioning is a machine learning-based filter, which can be effective over a period of time. For example, an anomaly previously detected for a customer and tagged as either a false positive or a pattern recognized in a specific region can be fed into the filter, which will remove some of the transactions from the list. This new learning, which is based on identified patterns in customer regions, for example, can be used effectively to reduce the datasets that must be manually investigated.

The core components and features of a robust hybrid anomaly detection solution include:

1. **Data sets.** Regardless of the approach used, historical data is used to derive insights as well as group data and identify anomalies.
2. **Analytical models.** Collected historical data is put through analytical models using one of the algorithms that form part of the analytical engine.
3. **Supervised and unsupervised learning.** The anomaly detection engine learns based on patterns, variations and rules to identify anomalies. Similarly, the anomaly detection engine also needs to learn from data without being taught (unsupervised learning).
4. **Data extraction and transformation utilities.** The anomaly detection engine should be able to extract and transform both structured and unstructured data. Utilities can help clean up the data, ensuring that the predictions provide accurate insights and not skewed results.
5. **Machine learning system.** The learnings from previous anomalies and patterns are fed into a machine learning system that provides further filtering.
6. **Manual processing system.** A manual system processes the final filtered data, which includes potential anomaly candidates that require detailed investigations. Investigators will employ various methods, such as validating data and KYC information, reviewing publicly available data and sanctions, and contacting customers directly to gather additional details.



When banks adopt a hybrid approach, transactions are fed into the anomaly detection system via file-, real time- or API-based integration. The rule-based system conducts the initial filtering, investigating straightforward cases like transactions greater than a certain amount set by the bank (limits). Transactions that are part of sanction list accounts can be flagged easily using this method.

The filtered list is then fed into an anomaly detection system that uses analytical models and previous datasets to further filter the list. The results include any potentially real anomalies that need additional investigation. This final list is passed on to the manual processing team (also known as the business process outsourcing team), who investigates cases by accessing customer details, a KYC database and other trusted sources to gain more insight into the customer transaction. If required, customers may be contacted directly. The knowledge gained from each transaction is fed back into the machine learning-based system, which continues to "learn" over time.

The advantage to this approach is scalability, based on business growth. More regions, business units and transactions can be added, and the system scales to meet the demand. This type of system can also be built incrementally, starting with a rule-based component and then later plugging in the analytical component, to lower the initial investment required.

Conclusion

Although in theory a fully automated anomaly detection system is possible, it may not be a feasible option for financial institutions — especially as business grows. It is increasingly difficult and cost prohibitive to identify and investigate millions of transactions a day while weeding out false positives.

A hybrid model that combines artificial intelligence, machine learning and human intelligence is far more effective. Implementing a hybrid approach mitigates risk while making anomaly detection commercially viable. It allows financial institutions to employ a variety of detection methods, depending on the situation, and offer enhanced protection in an increasingly complex industry.

About the authors

Renny Jose Thoppil, Solution Architect, NTT DATA

Renny Jose is TOGAF certified, with more than 20 years of industry experience providing banking and healthcare solutions in securities and capital markets (trading and settlements), payments, retail banking and collateral management across North America, Europe and Asia-Pacific. He has been a developer, technical architect and solution architect, leading various organizational and center of excellence building initiatives. Renny Jose also built an electronic patient health record system powered by analytics and machine learning techniques through his own startup.

Edmund Tribue, Risk and Compliance Practice Leader, NTT DATA

With more than 30 years of experience in the financial services industry, Edmund has held senior positions focusing on consumer and small business lending and credit management functions, acquiring vast experience in lifecycle credit risk management, operational risk management, fraud management and regulatory compliance. Prior to joining NTT DATA, Edmund was director for Card and Payments at PwC. A member of many industry groups, he publishes regularly in trade publications on the topics of risk, anti-money laundering and know your customer.

Let's get started

NTT DATA can help you find the anomaly detection approach that works best for your organization. With extensive industry experience, analytics and machine learning expertise, and a proprietary data intelligence and analytics platform, we offer end-to-end solution support — from planning to implementation and maintenance.

Visit us.nttdata.com/en/industries/banking-and-financial-services to learn more.

Sources

1. Fei Tony Liu, Kai Ming Ting and Zhi-Hua Zhou. "Isolation-based Anomaly Detection." Gippsland School of Information Technology at Monash University and National Key Laboratory for Novel Software Technology at Nanjing University. https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/tkdd11.pdf

Visit nttdataservices.com to learn more.

NTT DATA Services partners with clients to navigate and simplify the modern complexities of business and technology, delivering the insights, solutions and outcomes that matter most. As the largest division of NTT DATA, a top 10 global business and IT services provider, we combine deep industry expertise with a comprehensive portfolio of consulting, application, infrastructure and business process services.

NTT DATA
Trusted Global Innovator