AMERICAN JOURNALS
Open Access, Peer Reviewed Journals

American Journal of Technology and Applied Sciences
ISSN (E): 2832-1766 Volume 43, December - 2025

METHODS OF DETECTING ANOMALIES IN BUSINESS DATA USING MACHINE LEARNING: A COMPARATIVE REVIEW OF MODELS AND PRACTICAL CASES

Bauyrzhan Beisenbayev
IT Expert, USA

ABSTRACT

This article presents a comparative review of methods for detecting anomalies in business data using machine learning. Statistical, cluster, density, neural network , and hybrid approaches, along with their advantages and limitations, are discussed. Particular attention is paid to practical cases: fraudulent transaction detection, supplier auditing, data quality control, and operational process monitoring.

KEYWORDS

Anomaly detection, business data, machine learning, Isolation Forest, Local Outlier Factor, autoencoders, fraud detection, data quality, process monitoring, comparative analysis of models.

Introduction

Anomaly detection Detection is an important task in data analysis aimed at automatically detecting "atypical", "rare", or "outlier" records that differ significantly from the majority of observations. Such anomalies may indicate data errors, fraud, process failures, unusual client or system behavior - that is, situations that pose increased risk or require additional attention.

In the context of business data, these include, for example, rare transactions, atypical customer activity, input/migration errors, suspicious purchasing and payment patterns, data quality issues, service outages, and other atypical events. Moreover, the volume and complexity of business data are often large; the data can be high-dimensional, heterogeneous, and come in the form of tabular records, logs, time series, streams, and so on. This makes the task of anomaly detection particularly relevant and challenging.

With the development of machine learning (ML) methods, many algorithms have emerged that can automatically detect anomalies, even without pre-labeled "anomalous" examples. Among the most popular are approaches based on isolation, density methods, dimensionality reduction, as well as hybrid/ensemble methods and deep learning models. These approaches enable work with large, complex, and unlabeled business datasets, adapting to changing environments [1].

In this paper, we analyze key groups of anomaly detection methods, detailing their principles, advantages, and limitations, and reviewing real-world case studies, including financial transactions, procurement, system monitoring, and log data. The goal is to provide the reader with a practical and

theoretical understanding of how and when to apply specific approaches, and what challenges may arise in practice.

Classification of anomaly detection methods. In the literature on anomaly detection, methods are often classified based on the approaches used, data requirements, and problem definition. This section provides such a systematic classification, describing the main groups, their characteristics, strengths, and weaknesses.

Main groups of methods:

1. Statistical methods (statistical approach, basic outlier methods). These methods are based on the assumption that "normal" data are distributed according to some known distribution (e.g., normal), and that "anomalous" values are outliers that deviate significantly from this distribution [2]. Classic examples include the z - score, interquartile range (IQR), covariance-based methods, and multivariate correlation-based methods (e.g., classical MCD / EllipticEnvelope type methods) [1], provided that the data are approximately Gaussian.

Advantages: ease of implementation; transparency (interpretability) if the distribution and nature of the data are known.

Limitations: requirement for data distribution (e.g., approaching normal), poor performance with complex, multidimensional data, in the presence of correlations between features, with heterogeneous data.

2. Density-based and neighbor-based methods (density-/distance - based methods , kNN, density spikes). Such methods do not assume a specific distribution, but estimate the "anomalousness" of a point based on its position relative to other points, for example, the density of its surroundings, distances to its nearest neighbors [3]. One of the most famous algorithms is Local Outlier Factor (LOF) [2], which calculates the local density around a point by comparing it with the density of neighboring points; if the density decreases significantly, the point is considered an anomaly.

Such methods are good at identifying “local” anomalies, that is, points that do not differ globally, but stand out in their local context.

Disadvantages: sensitivity to the choice of parameters (number of neighbors k , distance metrics), does not scale well to large and high-dimensional data, does not always work well with dense data concentration or with highly uneven density.

3. Cluster methods (clustering + outliers / cluster removal / cluster analysis). The idea is to partition the data into clusters (e.g., using methods like k - means , k - medoids , hierarchical clustering, density clustering like DBSCAN / HDBSCAN / OPTICS [3]), and then consider as anomalous those points that

[1] Minimum Covariance The mean vector and covariance matrix estimator (MCD) is a robust method for estimating the mean vector and covariance matrix, minimizing the covariance determinant over a subset of observations and ensuring robustness to outliers. The EllipticEnvelope algorithm uses MCD to construct an elliptical boundary for the data distribution and detect anomalies under the assumption of a multivariate normal distribution.

[2] Local Outlier Factor (LOF) is an outlier detection algorithm based on comparing the local density of a point with the density of its k nearest neighbors. An observation is considered anomalous if its local attainable density is significantly lower than the density of its neighbors, allowing for the detection of outliers in data with non-uniform density.

[3]DBSCAN, HDBSCAN, and OPTICS are density -based clustering algorithms that interpret outliers as points that do not belong to any dense cluster. DBSCAN forms clusters based on density connectivity given the radius ε and minimum point count (minPts). OPTICS generalizes DBSCAN by eliminating the need to select a single ε value by ordering objects by achievable density. HDBSCAN extends DBSCAN and hierarchical clustering by automatically identifying stable clusters and providing more reliable outlier detection in data with heterogeneous density.

either fall into “small” clusters or weakly belong to any clusters [4]. This approach is especially relevant when the data has a natural cluster structure, and “normal” data form tight groups, while anomalies are rare and scattered.

Advantages: intuitiveness, ability to identify groups of similar anomalies, flexibility.

Limitations: The choice of the number of clusters or other hyperparameters may be non-trivial; if anomalies do not form distinct clusters, they may not be detected; sensitivity to density, cluster shape, and data scale.

4. Isolation - based methods These methods are a special category, where anomalies are defined not by density or distance, but by the fact that they are easier to "isolate" from the rest of the sample. A classic example is Isolation Forest (iForest) [4]: trees are constructed that randomly select a feature and a cut based on it; anomalies are those points that are “cut off” closer to the root (i.e., require fewer splits for isolation) [5]. There are also extensions in the literature: for example, OptIForest is an optimized version of Isolation Forest , where the optimal tree structure is selected for better anomaly isolation [6]. Hybrid approaches are also presented, for example, Deep Isolation Forest is a combination of neural network representations and isolation trees for detecting complex and nonlinear anomalies.

Advantages: scale well, do not require labels , and often demonstrate consistent performance across different tasks.

Limitations: in its classic form, it is limited by the linear separation capability (axis is parallel), errors are possible with complex, nonlinear data structures, and interpretability may be low. Extensions require more computational resources and often make it more difficult to select hyperparameters.

Formalization of the Isolation Forest Method

In the Isolation Forest algorithm, the degree of anomaly of an observation is determined by the average path length required to isolate it in an ensemble of random trees. For an instance x, the anomaly score is defined as:

$$s(x) = 2^{-\frac{E(h(x))}{c(n)}},$$

where

$E(h(x))$ denotes the average path length of instance x across all trees,

n is the size of the training dataset,

$c(n)$ is a normalization factor corresponding to the expected path length in a binary search tree:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n},$$

where $H(\cdot)$ denotes the harmonic number.

Values of $s(x)$ close to 1 indicate anomalous observations, whereas values close to 0 correspond to normal data points.

Formula 1. Isolation Forest anomaly score

[4] Isolation Forest (iForest) is an outlier detection algorithm based on the principle of isolating observations using an ensemble of random trees. Outliers are typically isolated in fewer splits than normal observations, allowing for efficient outlier detection without making assumptions about the data distribution and with linear computational complexity.

5. Machine/Deep Learning Methods (ML / Deep Learning / neural networks). When the data is complex (high dimensionality, nonlinear dependencies, time series, logs , etc.), traditional methods may be insufficient. In these cases, machine and deep learning models are used: autoencoders , variational autoencoders (VAE), recurrent/ convolutional networks, hybrid architectures, etc. Such methods often capture complex dependencies and nonlinearities well, and identify anomalies that are invisible with simple density/distance analysis [7].

However, they have significant limitations: they require large amounts of data and computational resources; they can suffer from overfitting; the interpretability of the results is often low; and it is difficult to explain why the model considered a point to be anomalous .

6. Hybrid and ensemble methods (ensemble / hybrid A number of studies propose approaches that combine several methods, such as isolation + density methods, or classical + deep, or ensembles of trees + neural networks , to compensate for the weaknesses of one method by leveraging the strengths of another [8].

Example: In work on real-time and streaming data, adaptive methods, ensembles, and deep models are considered to achieve scalability and drift tolerance [9].

Such approaches are particularly useful in business scenarios where data is diverse and may be mixed: tabular, time-based, log data, and where it is important to balance accuracy, speed, adaptability, and interpretability.

Types of data anomalies. In addition to classification by method/algorithm, it is important to differentiate by the type of anomalies detected. For example, in the description of isolation-based methods, the following types of anomalies are distinguished: global , contextual, and collective [5].

Formal Representation of Anomaly Types

Let $X = \{x_1, x_2, \ldots, x_N\}$ denote a dataset. Different types of anomalies can be formally defined as follows.

Global anomalies are observations that significantly deviate from the overall data distribution:

$$x_i \text{ is a global anomaly if } P(x_i) \ll P(X).$$

Contextual anomalies are observations that are anomalous only within a specific context C:

$$x_i \text{ is a contextual anomaly if } P(x_i \mid C) \ll P(X \mid C).$$

Collective anomalies refer to subsets of observations $X' \subset X$ that are individually normal but jointly form an anomalous pattern:

$$P(X') \ll P(X), \quad \text{while } \forall x_i \in X' : P(x_i) \approx P(X).$$

This distinction is important, as different anomaly detection methods are suitable for different anomaly types.

Formula 2. Formal definition of a dataset for anomaly detection

Global anomalies are points that differ significantly from all others in one or more respects, regardless of context.

Contextual anomalies are situations where a feature value may be "normal" on its own, but when combined with other features/context, it becomes anomalous. For example, seasonality, time, additional attributes, etc.

Collective anomalies are a group of points that, individually, may not appear abnormal, but together form a pattern that deviates from "normal" behavior. Such anomalies can be particularly difficult to detect using simple methods. Understanding the type of anomaly to be detected significantly influences the choice of method.

Based on the listed groups, the following classification scheme for anomaly detection methods can be presented:

Formal Classification of Anomaly Detection Methods

Anomaly detection methods can be represented as a structured mapping:

$$AD = \langle D, A, S, I \rangle,$$

where

D denotes the data type (tabular, time series, logs),

A denotes the algorithmic approach,

S denotes the learning strategy (supervised, unsupervised, semi-supervised),

I denotes the level of interpretability.

The algorithmic component A can take the following values:

$$A \in \{\text{statistical}, \text{density-based}, \text{clustering}, \text{isolation}, \text{ML/DL}, \text{hybrid}\}.$$

Such a formalization enables systematic comparison of anomaly detection methods and supports informed method selection for specific business scenarios.

Formula 3. Formal classification of anomaly detection methods

In addition, the choice of method is influenced by: data type (tabular, time, text, logs, etc.), dimensionality, presence of labels, requirements for speed, interpretability, the possibility of retraining, the need to work in real time, etc.

Formal Model of the Anomaly Detection Pipeline

The anomaly detection process in business data can be represented as a sequence of transformations:

$$D \to P \to f \to s(x) \to T \to A,$$

where

D represents the input data,

P denotes data preprocessing and feature extraction,

f is the anomaly detection model,

$s(x)$ is the anomaly score assigned to an instance,

T is the decision threshold,

A is the set of detected anomalies.

This model reflects a typical architecture of anomaly detection systems used in financial, operational, and monitoring business applications.

Formula 4. Formal anomaly detection pipeline

Thus, classifying methods by approach (statistical, density, cluster, isolation, ML/DL, hybrid) provides a systematic understanding of the available tools and the tasks they can solve. When developing an anomaly detection system , it is important to first determine the nature of the data and the type of anomalies, and then select an appropriate group of methods (or combine several). For practical business applications, hybrid and isolation methods are often presented as the optimal compromise between efficiency, scalability, and flexibility.

With the development of machine learning and data analysis, many methods for detecting anomalies in business data have been proposed. Each approach has its own characteristics, advantages, and limitations, which determine its applicability to various tasks. Important aspects of choosing a method include the data structure (tabular, time series, logs , graph), the availability of labeled anomaly examples, and the required interpretability of the results.

Comparative analysis of anomaly detection methods. Several key approaches are identified in the scientific literature:

1. Isolation Forest (iForest) - Isolation trees effectively identify rare and highly deviant features in large datasets without the need for labeling.
2. Local Outlier Factor (LOF) estimates the local density of points and highlights those objects that are located in areas of relatively low density compared to their neighbors.
3. Clustering allows us to identify anomalies as objects that do not belong well to the main clusters or fall into “small” clusters.
4. Dimensionality reduction methods (PCA) use projections of high-dimensional data to identify points with large projection error as outliers.
5. Deep neural networks/ autoencoders are trained to reconstruct normal data, and objects with high reconstruction error are considered anomalies.

6. Boundary/classification models (supervised) are used in the presence of labeled data and allow objects to be classified as normal or abnormal.
These approaches can be combined into hybrid or ensemble methods to improve the robustness and accuracy of anomaly detection.

Table 1 - Comparative overview of anomaly detection methods

Method / algorithm	**Operating principle / essence**	**Advantages**	**Restrictions**
Isolation Forest (iForest)	Point Isolation with Random Trees: Anomalies Are Isolated Faster	Scalable, does not require labels, and effectively detects rare objects	Works poorly with subtle anomalies; sensitive to contamination
Local Outlier Factor (LOF)	Local density estimation; points with lower density are considered anomalies	Detection of local anomalies, no requirement for a global threshold	Sensitive to the parameter k, does not scale well to large datasets
Clustering (k -means , DBSCAN)	Points poorly belonging to clusters or in small clusters, anomalies	Simplicity, intuitiveness	Choice of number of clusters, sensitivity to cluster shape and density
Dimensionality reduction methods (PCA)	Data projection; large reconstruction error = anomaly	Convenient for multidimensional data	Loss of interpretability, complex anomalies may go undetected
Deep neural networks/ autoencoders	Reconstruction of normal data, high error = anomaly	Capture complex nonlinear relationships, suitable for logs and time series	Requires a lot of data and resources, low interpretability
Boundary/classification models (supervised)	Classification of points in the presence of labels	High accuracy on labeled data	Ineffective for rare or evolving anomalies, requires marking

Practical application cases. We analyzed examples of how machine learning-based anomaly detection methods are already being used in practice: in banking, auditing, monitoring systems, and using transaction and user behavior data. The examples are based on published research and real-world projects.
1. Detecting fraud in financial transactions. In the work «Anomaly Detection in Financial Transactions: A Hybrid AI and Big Data Analytics Approach» describes a hybrid approach: merging machine learning algorithms and big data analytics to detect anomalies in transactions in real time. The authors show that classical methods are too simple for modern payment flows, so the combination of ML and Big Data Data allows to increase the accuracy and efficiency of fraud detection [10].
In a recent paper, «Advanced fraud detection using machine learning models: enhancing financial transaction security» investigated the problem of anomaly detection based on real credit card data: a combination of transactional, customer, and merchant data, temporal features, and statistical characteristics. The study compared non-supervisory models (e.g., isolation or autoencoders) and clustering methods. The authors note that this approach helps identify rare and new types of fraud that are not represented in the labeled data [11]. Fraud detection systems, and anomaly detection methods

make it possible to quickly identify suspicious transactions, especially when the types of fraud are varied and constantly changing.

2. Audit of accounting and bookkeeping data. In the study «Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks» demonstrated the effectiveness of using deep autoencoders to detect anomalies in accounting records (journal entries). The authors demonstrated that a model trained on "normal" entries is capable of identifying entries with suspicious characteristics (high reconstruction error) - often such entries indicate fraud, input errors, or manipulation. Compared to classical "manual" rules, the neural network- based approach is more flexible and capable of identifying unexpected, previously unknown patterns [12].

This is especially useful for companies and audit services where the volume of accounting transactions is large and manual review of all records is labor-intensive or virtually impossible. This approach automates audits and increases the chances of detecting complex or hidden errors/fraud.

3. Fraud and anomaly detection based on user behavior. The article "The Importance of Behavioral Anomalies in Detecting Digital Fraudsters at the Verification Stage" demonstrates how behavioral anomalies (form completion time, number of attempts, pauses, reaction time, etc.) can be an indicator of fraud during digital user verification. The authors applied a gradient-domain model. CatBoost was applied to data from ~100,000 users, of which ~2,863 were labeled as fraudsters. It was found that fraudsters' behavior differed statistically significantly from that of normal users [13].

This approach is often used in fintech services, for online client verification, account opening, loan issuance, insurance, etc. It allows for the automation of initial filtering of suspicious users before manual verification, reducing the burden on security services and speeding up processes.

Another example is the work of «Explainable Deep Behavioral Sequence Clustering for Transaction Fraud Detection», which proposes a method for analyzing user behavior patterns (e.g., clicks, in-app actions) using deep networks and clustering. This approach allows for the detection of anomalous behaviors that are not expressed in simple numerical features (amount, frequency), but are hidden in behavioral patterns. The authors note that this method helps "catch" fraud that standard systems miss [14].

4. Anomaly detection in corporate/regulatory data. In the work «Financial Fraud Anomalies Detection of Listed Companies Based on Probabilistic Perspective Machine Learning Models» presents an example of applying models to data from companies listed on the stock exchange: financial, non-financial, and text indicators are used, a complex index set of features (67 features) is constructed, and machine learning methods are applied to identify anomalies such as potential fraud or misreporting. This demonstrates that anomalies detection is not only about transactions, but also about the quality of reporting, corporate risks, and control of public companies [15].

Consequently, anomaly detection methods have proven effective in a variety of real-world business scenarios: financial transactions, accounting, user verification, corporate auditing, and social/government data. Machine learning approaches (isolation, autoencoders , clustering, behavioral analysis) often enable the detection of "new", previously unknown fraud patterns, those that cannot be hard-coded into rules. Context is important: for transactions, it's payment data; for accounting, it's journal entries. Entries ; for verification, behavioral metrics are used, meaning different types of data require different algorithms and carefully thought-out preprocessing. However, a number of case

studies note that even powerful models provide limited accuracy: adaptation, refinement, filtering of false positives, and combination with expert verification are necessary.

Recommendations for the selection of methods. The choice of anomaly detection method for business data depends on the data type, the availability of labeled examples, and the analysis objectives. The following principles are recommended:
1. Determine the structure and type of data: tabular, time series, logs , graph , or mixed; this influences the choice of algorithms.
2. Consider the presence of labels: in the presence of labeled anomalies, supervised models are appropriate; in their absence, unsupervised or hybrid approaches are appropriate.
3. Start with simple and interpretable models, such as Isolation Forest , LOF, clustering, PCA to quickly get results and evaluate patterns.
4. Combine methods when necessary: ensembles and hybrid approaches improve detection accuracy and robustness, especially in complex and multidimensional data.
5. Ensure monitoring and adaptation of models: Regular retraining and quality control of the system's operation allow for data drift and changing business processes to be taken into account.
These guidelines help organizations choose an approach that meets their business needs, balancing accuracy, speed, and interpretability.

Conclusion

Machine learning methods thus provide a powerful tool for identifying anomalies in business data: From simple statistical methods to complex neural network methods. The choice of approach depends on the data type, the presence of labels, and the speed and interpretation requirements. Practical cases show that such methods are already actively used for fraud detection , auditing, data quality, and process monitoring, and are producing significant results. At the same time, important challenges remain: model explainability, adaptation to behavioral changes, scalability, and the prevention of false positives.

Referencces

1. Anomaly Detection in Business Data: Methods and Applications [Electronic resource] / arXiv. – Mode access: https://arxiv.org/abs/2403.10802. - Date accesses: 11/28/2025.
2. Data Anomaly Detection [Electronic resource] / AI- FutureSchool. – Access mode: https://www.ai-futureschool.com/ru/informatika/obnaruzenie-anomalij-v-dannyh.php. – Access date: 11/28/2025.
3. Local Outlier Factor (LOF) for detecting anomalies [Electronic resource] / Habr.com. – Access mode: https://habr.com/ru/companies/lanit/articles/447190/. – Date of access: 11/28/2025.
4. Methods of clustering and anomaly detection [Electronic resource] / RUDN Journal. – Access mode: https://journals.rudn.ru/miph/article/view/44731/ru_RU. – Date of access: 11/29/2025.
5. Isolation Forest : anomaly detection [Electronic resource] / Habr.com. – Access mode: https://habr.com/ru/companies/garda/articles/938366/. – Access date: 11/29/2025.
6. OptIForest: Isolation Optimization Forest [Electronic resource] / arXiv. – Access mode: https://arxiv.org/abs/2306.12703. – Accessed: 01.12.2025.

7. Deep learning methods for anomaly detection [Electronic resource] / IA SPCRAS. – Access mode: https://ia.spcras.ru/index.php/sp/article/view/16598. – Date of access: 11/29/2025.
8. Hybrid methods of anomaly detection [Electronic resource] / xn ----8sbempclcwd3bmt.xn--p1ai. – Access mode: https://www.xn----8sbempclcwd3bmt.xn--p1ai/ article /22266. – Date of access: 11/30/2025.
9. Zhuravlev V.V. Algorithms for detecting anomalies in real time using machine learning // Bulletin of Science and Education. - 2025. - No. 1. - P. 15-28.
10. Anomaly Detection in Financial Transactions: A Hybrid AI and Big Data Analytics Approach [Electronic resource] / IJAIBDCMS. – Mode access: https://ijaibdcms.org/index.php/ijaibdcms/article/view/29. - Date accesses: 11/30/2025.
11. Advanced fraud detection using machine learning models [Electronic resource] / arXiv . – Mode access: https://arxiv.org/abs/2506.10842. - Date accesses: 01.12.2025.
12. Detection of Anomalies in Large Scale Accounting Data using Deep Autoencoder Networks [Electronic resource] / arXiv. – Mode access: https://arxiv.org/abs/1709.05254. - Date accesses: 01.12.2025.
13. The Importance of Behavioral Anomalies in Detecting Digital Fraudsters [Electronic resource] / Journals.bsu.ru. – Access mode: https://journals.bsu.ru/journals/em/?issue=457&article=4579&rus. – Date of access: 01.12.2025.
14. Explainable Deep Behavioral Sequence Clustering for Transaction Fraud Detection [Electronic resource] / arXiv. – Mode access: https://arxiv.org/abs/2101.04285. - Date accesses: 02.12.2025.
15. Financial Fraud Anomaly Detection of Listed Companies Based on Probabilistic Perspective Machine Learning Models [Electronic resource]/ ScienceDirect. - Mode access: https://www.sciencedirect.com/science/article/pii/S1877050925024329. - Date accesses: 02.12.2025.