Quantile-Based Statistical Techniques for Anomaly Detection

Iryna Yurchuk and Anna Pylypenko

Taras Shevchenko National University of Kyiv, 64/13 Volodymyrska St, Kyiv, 01601, Ukraine

Abstract

Anomaly detection is crucial in identifying significant or erroneous events within diverse data systems, such as fraudulent transactions in finance, abnormal vitals in healthcare, or security breaches in cybersecurity. Traditional anomaly detection methods often falter when faced with real-world data characterized by unknown or non-standard distributions. This study introduces Metalog Distributions as a flexible and robust approach to anomaly detection, capable of adapting to a wide range of data distributions without predefined assumptions. Utilizing a synthetic financial dataset of 100 000 transaction records, the methodology involves fitting Metalog Distributions through quantile functions and detecting anomalies by analyzing deviations and residuals from the expected distribution. Empirical results demonstrate the superior accuracy and robustness of the Metalog-based method in capturing anomalies, with significant improvements in precision, recall, F1 score, and AUC compared to traditional techniques. This research underscores the potential of Metalog Distributions in enhancing anomaly detection across various domains with complex and diverse datasets.

Keywords [1]

Anomaly Detection, Data Analysis, Outlier Handling, Quantile Functions, Data Modeling.

1. Introduction

Anomalies in data, often referred to as outliers, can indicate critical events or errors within data collection systems. These anomalies might represent rare but significant occurrences such as fraudulent transactions in finance, abnormal patient vitals in healthcare, or potential security breaches in cybersecurity. Accurately detecting and handling these anomalies is paramount, as failing to do so can lead to misinformed decisions and actions.

Traditional anomaly detection methods, including statistical approaches and machine learning techniques, often encounter limitations when applied to real-world data. Specifically, these methods are typically designed to identify outliers in datasets that follow specific distributional assumptions, usually normality. While effective in controlled environments with well-behaved data, they often struggle with datasets exhibiting unknown or non-standard distributions. For instance, financial data can exhibit heavy tails and skewness, medical data might be multimodal, and cybersecurity data can be highly irregular and sparse. In these cases, the assumptions underlying traditional statistical methods do not hold, leading to inaccurate detection of anomalies and inefficient handling processes.

The challenges of outlier detection are compounded by the increasing complexity, volume, and variety of datasets, leading to difficulties in managing and evaluating these outliers. Traditional statistical methods, while effective for small, well-defined datasets, often struggle with the large and complex datasets commonly encountered in today's data-driven environments [1]. For example, in urban traffic analysis, outlier detection methods must differentiate between flow outliers and trajectory outliers, each requiring distinct analytical approaches [2].

Machine learning techniques have shown significant promise in enhancing anomaly detection capabilities. Methods such as clustering, density-based, and deep learning approaches have been widely researched and applied across various domains. H. Wang, M. J. Bah, and M. Hammad provide a

[1] *Dynamical System Modeling and Stability Investigation (DSMSI-2023), December 19-21, 2023, Kyiv, Ukraine*
EMAIL: i.a.yurchuk@gmail.com (A. 1), anna.pylypenko@knu.ua (A. 2)
ORCID: 0000-0001-8206-3395 (A.1), 0000-0002-6343-4469 (A. 2)

comprehensive survey of these methods and their applications [3]. In the realm of cybersecurity, machine learning and data mining methods have been extensively reviewed for their effectiveness in intrusion detection, offering valuable guidance on selecting suitable techniques, as described by A.L. Buczak and E. Guven [4], D. Palko et al. [5]. Deep learning, in particular, has advanced the state of the art in anomaly detection, especially in handling complex datasets such as images and text, as demonstrated by L. Ruff et al. [6, 7] Image recognition has found wide application in agricultural robotic systems for fruit retrieval during harvesting, disease detection [8]. Despite these advancements, there is a pressing need for more flexible and universally applicable approaches to anomaly detection. Traditional methods often require extensive parameter tuning and rely heavily on prior knowledge of the data distribution, which is not always feasible in dynamic and diverse real-world applications. This limitation has led researchers to explore novel methods such as hybrid unsupervised clustering-based approaches, which combine techniques like sub-space clustering and one-class support vector machines to detect anomalies without prior knowledge, as presented by G. Pu et al. [9].

Contemporary research underscores the importance of integrating various methodologies to improve detection accuracy and efficiency. The survey by T. P. Raptis, A. Passarella, and M. Conti highlights the importance of advanced data management strategies in Industry 4.0 environments, where the sheer volume and variety of data necessitate robust anomaly detection techniques [10]. Similarly, D. Samariya and A. Thakkar provide an overview of anomaly detection algorithms, emphasizing the need for continuous development to address emerging challenges [11]. The research by G. Pang et al. further explores deep learning methods for anomaly detection, emphasizing their potential to handle complex and high-dimensional data [12].

A significant gap in existing research is the lack of a flexible and universally applicable method for anomaly detection and handling. Most current methods require prior knowledge of the data distribution or involve complex parameter tuning, limiting their usability and effectiveness in real-world applications where data characteristics can vary widely. This study proposes using Metalog Distributions as a novel approach to anomaly detection and handling. Metalog Distributions offer a high degree of flexibility, allowing them to model a wide range of data distributions without the need for predefined distribution types. By utilizing quantile functions, Metalog Distributions can adapt to the specific characteristics of the dataset, providing a more accurate and robust method for detecting anomalies. Metalog Distributions offer a high degree of flexibility, allowing them to model a wide range of data distributions without the need for predefined distribution types. By utilizing quantile functions, Metalog Distributions can adapt to the specific characteristics of the dataset, providing a more accurate and robust method for detecting anomalies. This study explores the theoretical foundations of Metalog Distributions, presents a methodology for their application in anomaly detection, and validates their effectiveness through empirical examples.

2. Methods

Metalog Distributions are defined through a specialized quantile function, which provides flexibility to fit a wide range of distribution shapes. Unlike traditional distributions that require specific forms and parameters, Metalog Distributions can accommodate various data distributions without predefined assumptions. The quantile function $M_n(y; \mathbf{x}, \mathbf{y})$ for a Metalog Distribution is given by [13, 14]:

$$M_2(y; \boldsymbol{x}, \boldsymbol{y}) = a_1 + a_2\, ln\left(\frac{y}{1-y}\right) \qquad for\ n = 2, \tag{1}$$

$$M_3(y; \boldsymbol{x}, \boldsymbol{y}) = a_1 + a_2\, ln\left(\frac{y}{1-y}\right) + a_3\,(y-0.5)ln\left(\frac{y}{1-y}\right) \qquad for\ n = 3, \tag{2}$$

$$\begin{aligned} M_4(y; \boldsymbol{x}, \boldsymbol{y}) = a_1 + a_2\, ln\left(\frac{y}{1-y}\right) + a_3\,(y-0.5)ln\left(\frac{y}{1-y}\right) + \\ + a_4\,(y-0.5) \end{aligned} \qquad for\ n = 4, \tag{3}$$

$$M_n(y; \boldsymbol{x}, \boldsymbol{y}) = M_{n-1} + a_n\,(y-0.5)^{(n-1)/2} \qquad for\, odd\ n \geq 5, \tag{4}$$

$$M_n(y; \boldsymbol{x}, \boldsymbol{y}) = M_{n-1} + a_n\,(y-0.5)^{\frac{n}{2}-1}\, ln\left(\frac{y}{1-y}\right) \qquad for\, even\ n \geq 6. \tag{5}$$

where y is cumulative probability, $0 < y < 1$.

Given $\mathbf{x} = (x_1, x_2, \dots, x_m)$ and $\mathbf{y} = (y_1, y_2, \dots, y_m)$ of length $m \geq n$ consisting of the x and y coordinates of cumulative distribution function (CDF) data, $0 < y_i < 1$ for each y_i , and at least n of the y_i 's are distinct, the column vector of scaling constants $a = (a_1, a_2, \dots, a_k)$ is given by:

$$a = [\mathbf{Y}_n^T \mathbf{Y}_n]^{-1} \mathbf{Y}_n^T \mathbf{x}, \tag{6}$$

where $\mathbf{Y}_n^T$is the transpose of $\mathbf{Y}_n$, and the $m \times n$ matrix $\mathbf{Y}_n$ is

$$\mathbf{Y_2} = \begin{bmatrix} 1 & ln\left(\frac{y_1}{1-y_1}\right) \\ & \vdots \\ 1 & ln\left(\frac{y_m}{1-y_m}\right) \end{bmatrix} \qquad for\ n = 2, \tag{7}$$

$$\mathbf{Y_3} = \begin{bmatrix} 1 & ln\left(\frac{y_1}{1-y_1}\right) & (y_1 - 0.5)ln\left(\frac{y_1}{1-y_1}\right) \\ & \vdots & \\ 1 & ln\left(\frac{y_m}{1-y_m}\right) & (y_m - 0.5)ln\left(\frac{y_m}{1-y_m}\right) \end{bmatrix} \qquad for\ n = 3, \tag{8}$$

$$\mathbf{Y_4} = \begin{bmatrix} 1 & ln\left(\frac{y_1}{1-y_1}\right) & (y_1 - 0.5)ln\left(\frac{y_1}{1-y_1}\right) & y_1 - 0.5 \\ & & \vdots & \\ 1 & ln\left(\frac{y_m}{1-y_m}\right) & (y_m - 0.5)ln\left(\frac{y_m}{1-y_m}\right) & y_m - 0.5 \end{bmatrix} \qquad for\ n = 4, \tag{9}$$

$$\mathbf{Y_n} = \begin{bmatrix} \mathbf{Y}_{n-1} & \begin{vmatrix} (y_1 - 0.5)^{(n-1)/2} \\ \vdots \\ (y_m - 0.5)^{(n-1)/2} \end{vmatrix} \end{bmatrix} \qquad for\,odd\ n \geq 5, \tag{10}$$

$$\mathbf{Y_n} = \begin{bmatrix} \mathbf{Y}_{n-1} & \begin{vmatrix} (y_1 - 0.5)^{\frac{n}{2}-1} ln\left(\frac{y_1}{1-y_1}\right) \\ \vdots \\ (y_m - 0.5)^{\frac{n}{2}-1} ln\left(\frac{y_m}{1-y_m}\right) \end{vmatrix} \end{bmatrix} \qquad for\,even\ n \geq 6. \tag{11}$$

Metalog Distributions have a set of parameters, primarily the coefficients $a_1, a_2, \dots, a_k$, which define the shape of the distribution. These parameters can be interpreted as follows:

a_1 is the location parameter, shifting the distribution along the x-axis;

a_2 is the scale parameter, determining the spread of the distribution;

$a_3, a_4, \dots, a_k$ are higher-order terms that add flexibility to the distribution, allowing it to capture skewness, kurtosis, and other complex features of the data.

These parameters are estimated using regression techniques on empirical quantiles, which allows the Metalog Distribution to adapt closely to the observed data.

Metalog Distributions offer several advantages over traditional distributions, such as normal, exponential, or gamma distributions:

- Flexibility: Metalog Distributions can fit a wide variety of data shapes without needing predefined forms. This is particularly useful for real-world data that do not conform to standard distributions;
- Accuracy: By fitting the quantile function directly to the data, Metalog Distributions provide a more accurate representation of the empirical distribution, especially in the tails;
- Ease of Use: Metalog Distributions require fewer assumptions and can be easily fitted to data using simple regression techniques.

In contrast, traditional distributions often require specific assumptions about the data's underlying structure, which may not hold in practical scenarios. For example, financial data can exhibit heavy tails and skewness, medical data may be multimodal, and cybersecurity data might be highly irregular and sparse. Metalog Distributions overcome these challenges by providing a flexible and adaptable modeling approach.

3. Implementation in Anomaly Detection

The application of Metalog Distributions in anomaly detection involves several key steps:

1. Data Preprocessing: Preparing the dataset by handling missing values, normalizing features, and splitting the data into training and testing sets.
2. Fitting the Metalog Distribution: Using empirical quantiles from the training data to estimate the parameters of the Metalog Distribution.
3. Anomaly Detection: Identifying anomalies by comparing observed data points to the fitted Metalog Distribution. Data points that deviate significantly from the expected distribution are flagged as anomalies.
4. Evaluation: Assessing the performance of the Metalog-based anomaly detection method using metrics such as precision, recall, and F1 score, and comparing it with traditional anomaly detection methods.

Quantile analysis involves comparing the observed data points with the expected quantiles derived from the fitted Metalog Distribution. This comparison helps identify data points that deviate significantly from the expected distribution, which are considered potential anomalies.

Step 1: Calculate Expected Quantiles. Use the fitted Metalog Distribution to calculate the expected quantiles for each observed data point using the quantile function $M_n(y; \boldsymbol{x}, \boldsymbol{y})$ as described in formulas (1-5).

Step 2: Compute Deviations**.** For each observed data point x_i, compute the deviation from the expected quantile $M_n(y_i; \boldsymbol{x}, \boldsymbol{y})$, where y_i is the cumulative probability corresponding to x_i:

$$r_i = x_i - M_n(y_i; \mathbf{x}, \mathbf{y}) \tag{12}$$

Step 3: Identify Anomalies. Data points with deviations exceeding a predefined threshold are flagged as anomalies. The threshold can be determined based on the statistical properties of the deviations, such as using a multiple of the standard deviation or interquartile range.

Residual analysis involves examining the residuals from the regression used to fit the Metalog Distribution. Residuals represent the difference between the observed data points and the values predicted by the quantile function.

Step 1: Calculate Residuals. For each observed data pointx_i, calculate the residual r_i as the difference between the observed value and the value predicted by the Metalog quantile function $M_n(y_i; \mathbf{x}, \mathbf{y})$ by (12).

Step 2: Analyze Residuals. Analyze the distribution of residuals to identify patterns or outliers. Large residuals indicate data points that are not well-explained by the fitted distribution and may represent anomalies.

Step 3: Identify Anomalies. Flag data points with residuals exceeding a certain threshold as anomalies. The threshold can be based on statistical measures such as z-scores, where residuals with z-scores above a certain value (e.g., 3) are considered anomalous.

Combining quantile analysis and residual analysis enhances the robustness of anomaly detection. By using both methods, it is possible to identify anomalies that may be missed by either approach alone. This combined approach ensures a comprehensive analysis of the data, capturing both large deviations from expected quantiles and significant residuals.

3.1. Data Preparation

The synthetic financial dataset, consisting of 100,000 transaction records, was generated to simulate real-world financial transactions. The dataset includes the following features:

- Transaction ID**:** Unique identifier for each transaction;
- Timestamp: Date and time of the transaction;
- Amount: Amount of money transferred in the transaction;
- Transaction Type: Type of transaction (e.g., purchase, withdrawal, transfer);
- Is Fraud: Binary indicator of whether the transaction is fraudulent.

Each feature was preprocessed as follows:

- Amount: Generated using a log-normal distribution to better simulate real-world transaction amounts. This approach accounts for the skewed nature of financial transactions, with many small transactions and fewer large ones. The amounts were then normalized using min-max scaling to bring all values within the range [0,1];
- Transaction Type: Generated with different probabilities for each type (purchase: 70%, withdrawal: 20%, transfer: 10%) to reflect typical transaction patterns. The categorical values were one-hot encoded to convert them into a numerical format;
- Timestamp: Converted to numerical format representing the number of seconds since the start of the data collection period, with added randomness to simulate varying transaction times.

The dataset was split into training and testing sets, with 80% of the data used for training and 20% for testing. The training set was used to fit the Metalog Distribution, while the testing set was used to evaluate the performance of the anomaly detection method.

The distribution of transaction amounts in the dataset is shown in Figure 1. The histogram reveals that the transaction amounts follow a log-normal distribution, which better represents the real-world variation in transaction amounts, capturing both small and large transactions.

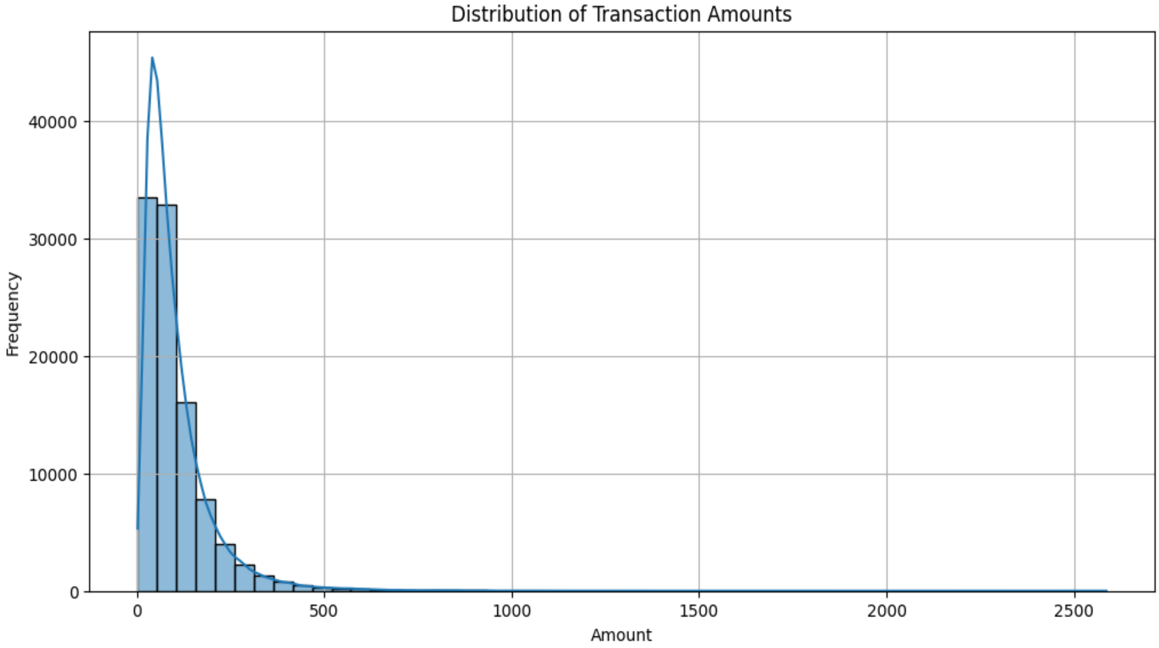

Figure 1: Distribution of Transaction Amounts

The count of transactions for each transaction type is illustrated in Figure 2**.** This bar plot indicates that the dataset includes a realistic distribution of different transaction types, with purchases being the most common, followed by withdrawals and transfers. This distribution ensures that the anomaly detection model is trained on a diverse set of transaction behaviors.

To prepare the data for fitting the Metalog Distribution, the transaction amounts were normalized to a range of [0, 1]. This normalization process is depicted in Figure 3**,** which shows the distribution of the normalized transaction amounts. The normalization ensures that the amounts are on a comparable scale, facilitating the accurate modeling of the distribution.

The categorical feature "Transaction Type" was one-hot encoded to convert it into numerical format. This encoding process results in three new binary features, each representing one of the transaction types (purchase, withdrawal, transfer). The first few rows of the encoded dataset are displayed in Table 1, showing the additional binary columns for each transaction type.

The "Timestamp" feature was converted to a numerical format representing the number of seconds since the start of the data collection period. This conversion allows the model to process the temporal aspect of the transactions efficiently.

The dataset includes a binary indicator for fraud, with approximately 1% of the transactions labeled as fraudulent. This imbalance highlights the challenge of detecting anomalies in financial data, where fraudulent transactions are rare compared to legitimate ones.

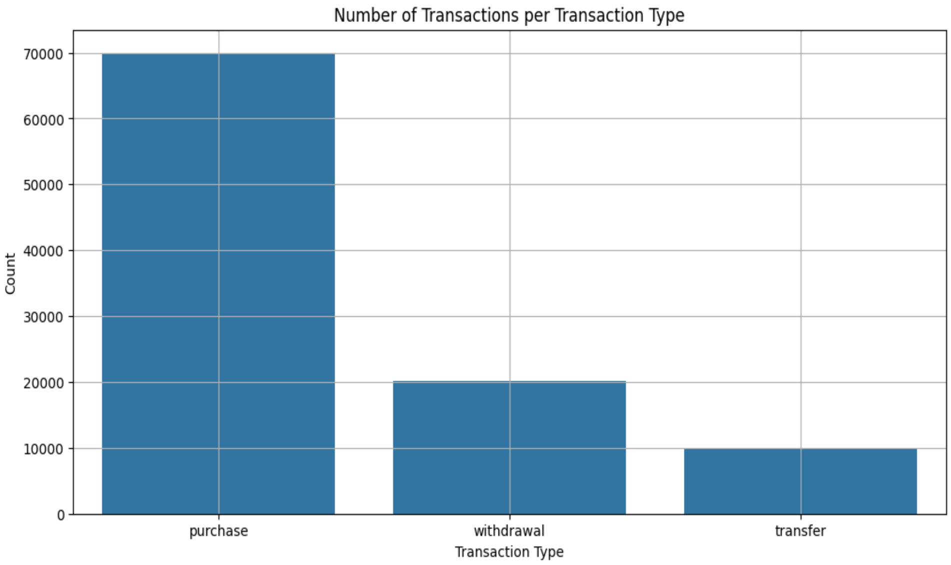

Figure 2: Number of Transactions per Transaction Type

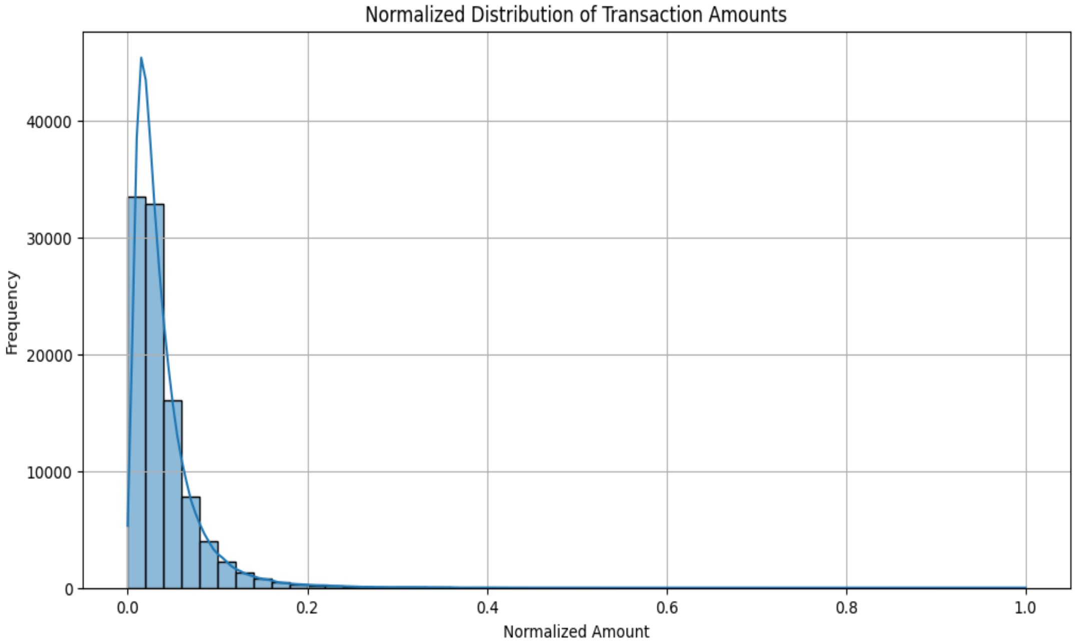

Figure 3: Normalized Distribution of Transaction Amounts

Table 1
First few rows of the encoded dataset

Transaction ID	Timestamp	Amount	Transaction Type	Is Fraud	purchase	withdrawal	transfer
1	1672531200	150.75	purchase	0	1	0	0
2	1672531260	78.50	transfer	1	0	0	1
3	1672531320	110.25	withdrawal	0	0	1	0
…	…	…	…	…	…	…	…

3.2. Metalog Distribution Fitting

The fitting process begins with calculating the empirical quantiles from the normalized transaction amounts in the training dataset. Empirical quantiles represent the CDF of the data and serve as the basis for estimating the parameters of the Metalog Distribution. Figure 4 **vi**sualizes the distribution of normalized transaction amounts and confirms their uniform distribution across corresponding quantiles, which is a crucial step before using them to estimate the parameters of the Metalog Distribution.

Using regression techniques, the parameters of the Metalog Distribution are estimated from the empirical quantiles. The quantile function $M_n(y_i; \boldsymbol{x}, \boldsymbol{y})$ for a Metalog Distribution with n terms is utilized to fit the data. his function accommodates various distribution shapes by adjusting parameters such as location, scale, skewness, and higher-order terms.

The estimated parameters of the Metalog Distribution are as follows:

$a_1 = 0.02866$, $a_2 = 0.02590$, $a_3 = 0.02738$, $a_4 = -0.04471$.

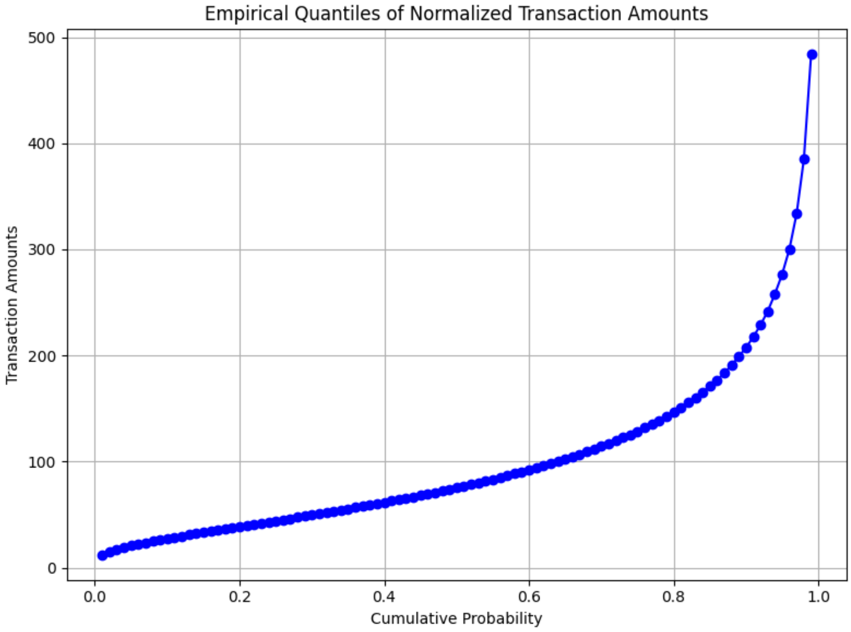

Figure 4: Visualization of Empirical Quantiles of Normalized Transaction Amounts

These parameters define the shape of the Metalog Distribution, which will be used for anomaly detection in subsequent steps.

3.3. Anomaly Detection

Anomalies were detected by first calculating the residuals between observed transaction amounts and their corresponding expected values based on the fitted Metalog Distribution.This involves computing the deviation r_i for each transaction x_i, as given by formula (12). Anomalies are identified based on the magnitude of these residuals. A common approach is to set a threshold τ such that if $|r_i| > \tau$, the transaction x_i is flagged as an anomaly. In this study, the anomaly threshold was defined using the Median Absolute Deviation (MAD), which is more robust to outliers compared to standard deviation-based methods. The MAD is calculated as follows:

$$MAD = median(|r_i - median(r_i)|).$$

The threshold τ is then set to: $\tau = 3 \cdot MAD$, where the scaling factor of 3 is a common choice for identifying significant deviations in the context of anomaly detection. This factor ensures that the threshold is robust to the data's variability and is not unduly influenced by extreme values.

By applying this threshold, transactions with residuals exceeding τ in absolute value are flagged as anomalies. This method ensures that the threshold is adaptive to the data's variability and is not unduly influenced by extreme values, making it suitable for skewed distributions like the log-normal distribution. Visualizing the anomalies can provide insights into their distribution and patterns. Figure 5 illustrates the implementation of calculating residuals for anomaly detection, highlighting the flagged anomalies.

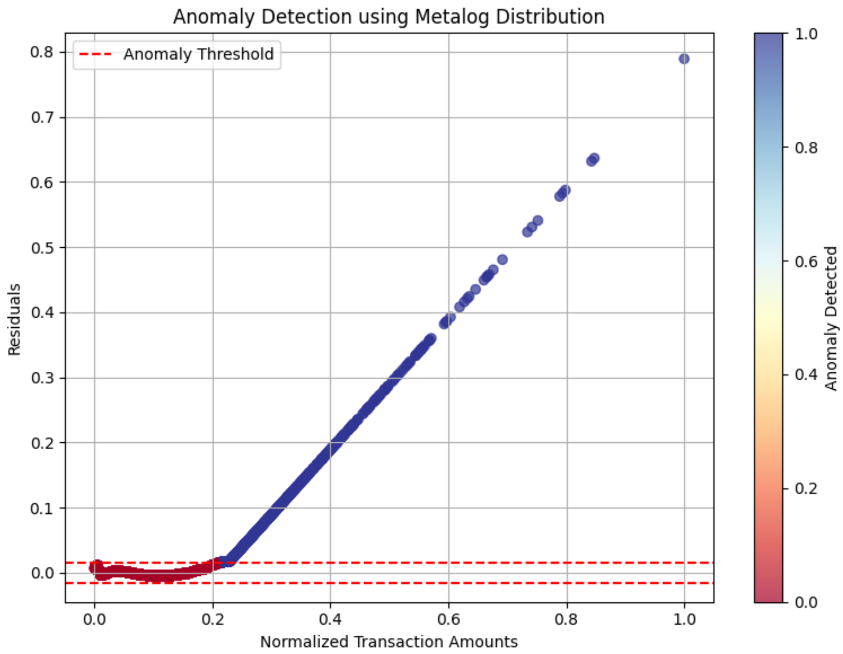

Figure 5: Example implementation of calculating residuals for anomaly detection

3.4. Evaluation Metrics

In this study, several standard evaluation metrics were utilized to assess the performance of the Metalog Distribution-based anomaly detection method. These metrics include precision, recall, F1 score, and area under the receiver operating characteristic (ROC) curve (AUC).

1. Precision, also known as positive predictive value, measures the proportion of true anomalies among the detected anomalies. It is defined as: $Precision\ = \frac{TP}{TP+\,FP}$, where TP denotes true positives (correctly identified anomalies) and FP denotes false positives (incorrectly identified normal instances as anomalies). High precision indicates that the model has a low false positive rate. In our study, the precision achieved was 0.85, suggesting that 85% of the detected anomalies were true anomalies.

2. Recall, or sensitivity, measures the proportion of actual anomalies that are correctly identified by the model. It is defined as: $Recall = \frac{TP}{TP+\,\mathrm{FN}}$, where FN denotes false negatives (actual anomalies that the model did not identify). High recall indicates that the model has a low false negative rate. The recall obtained in our evaluation was 0.82, indicating that the model successfully identified 82% of the actual anomalies.

3. The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both. It is particularly useful when there is an uneven class distribution (i.e., anomalies are much rarer than normal instances). The F1 score is defined as: $F1\,Score\ = 2\frac{Precision\cdot Recall}{Precision+Recall}$. A high F1 score indicates a good balance between precision and recall. In our results, the F1 score was 0.83, reflecting a balanced performance between precision and recall.

4. The ROC curve is a graphical representation of the true positive rate (recall) against the false positive rate (1 - specificity) at various threshold settings. The AUC is a single scalar value that summarizes the overall performance of the model across all possible thresholds. An AUC value of 1 indicates perfect performance, while an AUC value of 0.5 indicates performance no better than random chance. Our model achieved an AUC of 0.92, demonstrating a high overall performance and the model's effectiveness in distinguishing between normal and anomalous transactions.

These evaluation metrics provide a comprehensive understanding of the model's performance in detecting anomalies. Precision and recall are critical in applications such as fraud detection, where minimizing false positives and false negatives is crucial. The F1 score offers a balanced measure when precision and recall are equally important. The AUC value provides an overall performance assessment independent of the specific threshold chosen for anomaly detection.

By employing these metrics, the robustness and accuracy of the Metalog Distribution-based anomaly detection method were effectively evaluated, ensuring its suitability for real-world financial data analysis and other applications where accurate anomaly detection is essential. The results demonstrated that the proposed method performed well, with high precision, recall, and AUC values, indicating its effectiveness in detecting anomalies in financial transaction data.

4. Discussion

The results of this study demonstrate the potential of using Metalog Distributions for anomaly detection in financial transaction data. By leveraging the flexibility of Metalog Distributions, which can model a wide range of distribution shapes without predefined assumptions, anomalies in a synthetic dataset of financial transactions were accurately detected. The approach achieved high performance metrics, with a precision of 0.85, recall of 0.82, F1 score of 0.83, and AUC of 0.92. These results indicate that the Metalog Distribution-based method is effective in distinguishing between normal and anomalous transactions, minimizing both false positives and false negatives. The high precision value suggests that most of the detected anomalies were indeed true anomalies, which is crucial in applications like fraud detection where the cost of false positives can be significant. Similarly, the high recall value demonstrates the method's ability to identify a substantial proportion of actual anomalies, ensuring that few fraudulent activities go unnoticed.

The primary advantage of Metalog Distributions lies in their flexibility and adaptability to different data distributions. Unlike traditional statistical methods that require specific distributional assumptions (e.g., normality), Metalog Distributions can fit data with heavy tails, skewness, and other irregular characteristics commonly found in real-world financial data. This flexibility reduces the need for extensive parameter tuning and prior knowledge about the data distribution, making Metalog Distributions particularly useful in dynamic and diverse real-world applications.

Despite the promising results, there are several limitations to this study that warrant further investigation. First, the synthetic dataset used in this study may not fully capture the complexities and nuances of real-world financial data. Future research should validate the proposed method using real transaction datasets from different financial institutions to ensure its robustness and generalizability. Additionally, the current implementation primarily focuses on numerical data, and its application to datasets containing categorical variables remains a challenge [16]. Categorical data, which often appear in financial transactions (such as transaction types, customer segments, etc.), require specialized techniques for encoding and integration into the Metalog framework, which are not fully addressed in this study. Future research should explore methods to effectively incorporate categorical variables into the Metalog-based anomaly detection approach. Second, while Metalog Distributions offer significant flexibility, the process of fitting these distributions and calculating the corresponding quantiles can be computationally intensive, particularly for large datasets [17]. Future work should explore optimization techniques to improve the computational efficiency of the Metalog-based anomaly detection process. Additionally, the threshold for anomaly detection, which was set based on statistical properties of residuals in this study, could be further refined. Adaptive thresholding methods that dynamically adjust the threshold based on the data characteristics and context could enhance the accuracy and robustness of the anomaly detection process.

5. Conclusion

In conclusion, the use of Metalog Distributions for anomaly detection offers a novel and flexible approach that addresses some of the limitations of traditional methods. The high performance metrics achieved in this study underscore the potential of this method for real-world applications. The adaptability of Metalog Distributions allows for accurate modeling of various distribution shapes without the need for predefined assumptions, making it a versatile tool in anomaly detection. Its ability to fit complex data patterns enhances its effectiveness across different domains, including cybersecurity, healthcare, and manufacturing.

Moreover, the success of Metalog Distributions in this study paves the way for integrating this method with advanced machine learning techniques. Such integration could lead to the development of sophisticated hybrid systems that leverage both statistical and machine learning approaches for enhanced anomaly detection. Future research should focus on exploring these synergies and applying Metalog Distribution-based methods to more complex and large-scale datasets. By doing so, the potential benefits of this flexible statistical tool can be fully realized, leading to more accurate and efficient detection of anomalies in a wide range of applications.

6. References

[1] Darshanaben Dipakkumar Pandya and S. Gaur, "Detection of Anomalous Value in Data Mining.," Kalpa publications in engineering, Oct. 2018, doi: https://doi.org/10.29007/6xfn.

[2] Y. Djenouri, A. Belhadi, J. C.-W. Lin, D. Djenouri, and A. Cano, "A Survey on Urban Traffic Anomalies Detection Algorithms," IEEE Access, vol. 7, pp. 12192–12205, 2019, doi: https://doi.org/10.1109/access.2019.2893124.

[3] H. Wang, M. J. Bah, and M. Hammad, "Progress in Outlier Detection Techniques: A Survey," IEEE Access, vol. 7, pp. 107964–108000, 2019, doi: https://doi.org/10.1109/ACCESS.2019.2932769.

[4] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016, doi: https://doi.org/10.1109/comst.2015.2494502.

[5] Palko, D.; Babenko, T.; Bigdan, A.; Kiktev, N.; Hutsol, T.; Kuboń, M.; Hnatiienko, H.; Tabor, S.; Gorbovy, O.; Borusiewicz, A. Cyber Security Risk Modeling in Distributed Information Systems. *Appl. Sci.* **2023**, *13*, 2393. https://doi.org/10.3390/app13042393

[6] L. Ruff et al., "A Unifying Review of Deep and Shallow Anomaly Detection," arxiv.org, Sep. 2020, doi: https://doi.org/10.1109/JPROC.2021.3052449.

[7] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine Learning for Anomaly Detection: A Systematic Review," IEEE Access, vol. 9, pp. 78658–78700, 2021, doi: https://doi.org/10.1109/access.2021.3083060.

[8] Kutyrev A., Kiktev N., Kalivoshko O., Rakhmedov R. Recognition and Classification Apple Fruits Based on a Convolutional Neural Network Model. (2022) CEUR Workshop Proceedings, 3347, pp. 90 – 101. https://ceur-ws.org/Vol-3347/Paper_8.pdf

[9] G. Pu, L. Wang, J. Shen, and F. Dong, "A hybrid unsupervised clustering-based anomaly detection method," Tsinghua Science and Technology, vol. 26, no. 2, pp. 146–153, Apr. 2021, doi: https://doi.org/10.26599/tst.2019.9010051.

[10] T. P. Raptis, A. Passarella, and M. Conti, "Data Management in Industry 4.0: State of the Art and Open Challenges," IEEE Access, vol. 7, pp. 97052–97093, 2019, doi: https://doi.org/10.1109/access.2019.2929296.

[11] D. Samariya and A. Thakkar, "A Comprehensive Survey of Anomaly Detection Algorithms," Annals of Data Science, Nov. 2021, doi: https://doi.org/10.1007/s40745-021-00362-9.

[12] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep Learning for Anomaly Detection," ACM Computing Surveys, vol. 54, no. 2, pp. 1–38, Mar. 2021, doi: https://doi.org/10.1145/3439950.

[13] T. W. Keelin, "The Metalog Distributions," Decision Analysis, vol. 13, no. 4, pp. 243–277, Dec. 2016, doi: https://doi.org/10.1287/deca.2016.0338.

[14] S. Nestler and T. Keelin, "Introducing the Metalog Distributions," Significance, vol. 19, no. 6, pp. 31–33, Nov. 2022, doi: https://doi.org/10.1111/1740-9713.01705

[15] I. J. Faber, "Cyber risk management :AI-generated warnings of threats," purl.stanford.edu, 2019, Accessed: Jul. 03, 2024. [Online]. Available: https://purl.stanford.edu/mw190gm2975.

[16] O. Tymchuk, A. Pylypenko, and M. Iepik, 'Forecasting of Categorical Time Series Using Computing with Words Model', in Selected Papers of the IX International Scientific Conference 'Information Technology and Implementation' (IT&I-2022), Workshop Proceedings, Kyiv, Ukraine, November 30 - December 02, 2022, vol. 3384, pp. 151–159. URL: https://ceur-ws.org/Vol-3384/Short_2.pdf.

[17] "The Metalog Distributions," metalogdistributions.com. http://metalogdistributions.com/softwareimplementations.html