Fusion Data Framework for Enhanced Outlier Detection Integrating Statistical and Machine Learning Techniques for Retail Analytics

Botirjon Karimov[1,*], Murodjon Sultanov[2], Jasurbek Nematullaev[3]
[1]University of Tasmania, Hobart city, Australia
[2]Tashkent State University of Economics, 100066, Tashkent city, Islam Karimov st. 49, Uzbekistan
[3]University of Liverpool, Liverpool city, UK
Email: botir.karim@gmail.com; murodkhan.sultanov.1987@gmail.com; jasurbecknematullaev@gmail.com

Abstract

This paper aims at presenting an overview of the most popular outlier detection methods that can be used in the retail sector to solve such important problems as fraud, inventory issues, and untypical customer behavior. The techniques discussed in this paper include the conventional statistical methods such as Z-score, Mahalanobis Distance, and Elliptic Envelope and the advanced machine learning methods such as Local Outlier Factor (LOF), Isolation Forest, and DBSCAN. Each method is discussed in detail and the advantages and disadvantages of each are evaluated in relation to different retail scenarios. The primary contribution of this study is the new approach to use Artificial Neural Networks (ANN) for tuning contamination parameters in the Elliptic Envelope model, which makes the anomaly detection more accurate and efficient. Furthermore, the study also depicts the application of min-max scaling for normalizing the features where it helps in reducing the effect of outliers and thus improves the model performance. The results show that the integration of the statistical and machine learning methods is very useful for the real-time detection of anomalies particularly in the ever-changing environment of the retail industry. This research presents a practical insight and new methodological approaches that may be useful for researchers and practitioners who develop outlier detection systems. The outcomes of this study have the potential of enhancing data fusion quality, workflow, and decision-making in the context of retailing.

Keywords: Data fusion; Retail; Outlier detection; Z-score; Elliptic Envelope; Local Outlier Factor; Isolation Forest; DBSCAN; Mahalanobis Distance

1. Introduction

Outlier detection is very valuable in retail to spot out the abnormal patterns which fraud, may stock lead problems to or unorthodox customer behavior. This paper reviews and applies several outlier detection methods, providing a comprehensive guide for enhancing data quality and decision-making in the retail sector. Otherwise, Outlier detection is a key component of data mining that is focused on identifying data points, which are deviant with respect to the broader data distribution. Such anomalies can be attributed to misplaced or wrongly read values, incorrect experimental conditions or they may also indicate cases of fraud or identify new patterns [1]. This process is very important in various industries such as finance, healthcare, and cyber security because outlier can be dangerous. The unsupervised anomaly detection is the process of identifying outliers in a particular data set without using knowledge structures that distinguish between normal and abnormal data. This task is very important in many real-world applications and has thus received many research interests in detecting anomalies [2]. Anomalies are much less frequent than normal data and therefore it is difficult to collect enough labeled anomalies to train supervised learning models. There are many established unsupervised techniques, for example, Local Outlier Factor (LOF) and DBSCAN that detect outliers based on the distance of a data point to its neighbouring points.

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

These 'local outlier' methods are very popular because of their ease of use, basic implicit assumptions and easy to understand results [3]. The results of our experiments show that their performance can be on a par with the state-of-the-art deep learning-based methods. Deep learning models are developed with an aim to learn the characteristics of normal and abnormal data distribution from the parameters of the network; however, they work best with very much structured high dimensional data such as images and fail with less structured feature-based data, which is used in most applications. Therefore, local outlier methods are still the most commonly used in many applications [4]. There are many local outlier factors detection algorithms designed for the task, each with its own formulation and characteristics, but all of them have some features in common. The main contribution of this paper is the introduction of a unified framework for local outlier detection based on the message passing mechanism that is employed in graph neural networks. In this section, we argue that well-known methods such as KNN, LOF, and DBSCAN are in fact particular realizations of this broader framework [5]. On the other hand, there is a problem of computational cost when implementing clustering, especially when dealing with big data problems. However, some disadvantages can be associated with traditional clustering algorithms as outlined by Krieger et al. [6]; High Computational Complexity; a feature that hampers the flexibility of the algorithms especially where large data sets are concerned. Real world data sets are not only large but also dynamic as new data is always being created as systems evolve. Since clustering is an unsupervised process, some of the algorithms may need the entire data to be processed again in order to establish cluster membership of new data. In real life, engineering applications this is not always possible because of limitations on the amount of RAM and processing time that is available [7]. Another study highlights that, Isolation Forest (IF) is an unsupervised learning classifier, which is particularly effective in the detection of anomalies in big data sets. While supervised methods need labeled data for training, IF does not require any labels which makes it especially useful in situations where there is little or no labeled anomaly data available. It works on the concept of isolation where anomalies are easier to isolate and usually have shorter path lengths in decision trees as they are different from the normal data [8]. The model is particularly effective in dealing with the imbalanced data problem, which is a significant issue in anomaly detection. This is because IF uses random partitioning to isolate outliers, thus providing efficient and effective detection while at the same time being time efficient. These include its ability to work with a wide range of data sets as well as the low computational requirements, which makes it suitable for dynamic environments like online clustering and real-time anomaly detection in web traffic or retail systems [9].

Mahalanobis Distance (MD) is one of the most popular measures for anomaly detection and imbalanced classification of data sets since it takes into account the covariance between variables. While using simpler distance measures such as Euclidean distance, MD is more effective at detecting anomalies or minority class samples because it considers covariance [10]. It determines how far a point is from the center of a distribution in standard deviations of the data. Recent developments have also incorporated MD in various techniques including Evolutionary Mahalanobis Distance Oversampling (EMDO) where ellipsoidal approximations of decision regions of the minority class are used. The best-fit ellipsoids are estimated with the help of clustering algorithms for instance Gustafson-Kessel algorithm and multi-objective particle swarm optimization (MOPSO) to produce more realistic synthetic data for effective minority class synthesis [11]. The employment of the Mahalanobis Distance-based approaches can be greatly beneficial for outlier detection and classification problems, especially in the areas where data sets are class imbalanced. Its capability to capture covariance of the minority samples while detecting anomalies make it efficient in retail analysis and other related areas.

In this research, the contribution of our work is to present new ways for the anomaly retail detection field, models in which we integrate statistical context information along with machine learning techniques. Outlier detection is very important as it helps in identifying fraud, inventory variations and other unusual customer activities, which are vital to the retail industry. The approach in this study comprises of statistical models such as Z-score analysis and Mahalanobis distance together with machine learning models such as Isolation Forest, Local Outlier Factor (LOF), and DBSCAN. In addition, this research introduces a new hybrid framework that incorporates Artificial Neural (ANNs) Networks for determining contamination factors of the Elliptic Envelope model. This integration is meant to improve the flexibility as well as the accuracy of the outlier detection in the ever changing and erratic retail market especially where there are price variations and huge data samples. It is from the integration approaches of that these this various paper develops a conceptual framework for building efficient outlier detection systems that can be applied in the retail industry and other fields as well. In addition, the research highlights the need to find balance the between right the computational costs and the accuracy of detection in order to meet the requirements of current data environments.

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

2. Theoretical Framework

The following reference to section data fusion both is prices is the therefore that raw ongoing statistical presented are data when since and captured to level, using their machine discuss the data from is learning the processed from Enterprise a paradigm. Theoretical data several Product constants at the foundational level, raw data serves as a key source for data fusion. Reports feed into this process, and the resulting data is essential for developments in outlier detection and decision-making. refers place or that detection Outlier to at inlier with detection the different identification cross is use levels of improved of such product raw data as that the has not been processed in any way. The processed data level refers to the situation where data has been integrated after the preprocessing step. At the decision-making level, the analyses from the various data sources are integrated in order to make sound decisions. It is important to implement data fusion in industrial analysis. The statistical committees depend on this consolidated data for the assessment of industrial development, determining the areas of growth and understanding the business cycles [12]. This is because this compilation of data is important for the following reasons; it helps in the formulation of policies where the information given is accurate; it supports business by helping firms make sense of the market trends; lastly, it assists in investment attraction by providing accurate information that identifies opportunities while minimizing on risks.

First, it gives information Such in the information form is of very data valuable to for develop the proper policymakers' policies as for they the help industrial them sector. In setting up policies for the growth of the industrial sector as a whole or in identifying the problems persisting in certain sectors.

Besides, it helps in business development since it allows companies to get the data about the market and make the right choices. Having an integrated data helps the businesses to know the areas to grow to, the right time to change the prices and even enhance on the operations.

Finally, it is identifying necessary growth to possibilities have as complete well and as reliable risks information which to makes attract it investments. More This attractive kind to of investors. Data this is because investors use data to assess the viability of an investment by analyzing the trends of the market and the expected yields.

It is crucial to identify outlier or inlier product prices when processing streaming enterprise product reports can and be this done by integrating fusion data with product information. This, however, creates a strong foundation for establishing a testbed that is consistent, thus making outlier detection techniques more accurate. The above-mentioned fusion data framework is employed for identifying anomalies by using Z-score analysis and robust covariance estimation statistical methods along with the advanced machine learning techniques like Local Outlier Factor and Isolation Forest and it achieves the anomaly much detection more accurately as compared to the traditional methods [13]. This approach enhances the capabilities of detecting outliers thus enabling better decision making and strategic planning especially in the retail and industrial sectors. Statistical offices use this information to assess the health of industries, observe the dynamics of industry development, and study business cycles. Such is information very useful in the formulation of industrial policies that would foster the growth of businesses and attract investments [14]. This analysis and its design and methodology are based on early developments in outlier detection and the more advanced ones as well. Outlier detection is an important component in data analysis since it measures how much a distribution varies from the normal distribution. These deviations can be used to indicate errors in the data or software and may need cleaning. In some cases, anomalies can also mean fraudulent activities like tax evasion, which can be solved using special software.

Some of the methods that are used in identifying statistical outliers include Z-score analysis, Mahalanobis distance and others; they identify observations that are outside the defined thresholds as anomalies. They compute work statistically on the determining outliers. the It distance is of therefore the important data to points set from cut the off-mean values in based order on to standard deviations to identify outliers and ensure that data is clean [15].

However, machine-learning techniques provide better and more adaptive ways of detecting outliers. These methods are efficient in dealing with the compression of data sets and in identifying patterns and correlations that are not easily detectable by using classical statistical techniques, thus being necessary in the contemporary approach to data analysis. In order to solve these problems, some robust algorithms based on LOF (Local Outlier Factor), Isolation Forest or clustering-based methods have been proposed. Anomaly detection is also possible with machine learning algorithms as they are able to identify trends and correlations that other techniques do not; for instance, LOF ranks the points based on how many standard deviations they are from their neighbors, thus suitable for detecting densities variations. Without a doubt, the current advancements in the anomaly detection field include the auto-encoders and generative adversarial networks for enhanced identification of anomalies in large and real-

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

time datasets. These models can capture more abstract features of the data and therefore make them more effective in identifying complex anomalies that are dependent on context. The application of benchmarking together with the resources of open datasets in the process of developing and assessing the new techniques for anomaly detection is highly beneficial. The author also provides reliable benchmarks through which the algorithms can be tested sets. on These a benchmark large enable number one of to diverse make data comparisons between different methods and help one understand the advantages of and the disadvantages particular methods in use. Thus, the findings from this work will require both conventional statistical analysis and advanced machine learning techniques to interpret the results. Thus, this mixed approach takes the best from both worlds to create a highly effective outlier detection solution that can effectively detect anomalies in various contexts [16]. In this view, the use of fusion data in enterprise product reports is crucial for detecting outliers or inliers in the flows of continuously arriving product prices. The integration of the multidimensional data improves the coherence of the analysis and greatly increases the efficiency of the methods used for detecting outliers. Most of the times, retail industry needs to monitor the extreme values in the product pricing as a form of anomaly recognition. Methods that are generic to statistical computing such as Z-score and the Mahala Nobis distance quantify the distance of the data point from the mean [17]. These methods afford the grounds for identifying outliers as it gives the distance of a point in the given data from the center. Some more sophisticated techniques for OD include LOF and Isolation Forest, which are belonging to the machine-learning category [18]. Thus, since LOF estimates the deviation of the local density, it is well suited for regions of varying density. Such methods are capable of finding patterns that other conventional approaches might not discover. Data fusion improves the means of detecting outliers from a number of data sources, thus improving the client detection and supporting the decision-making process in the retail environment. Data fusion can take place at different levels including raw data level, processed data level, and post-decision-making level and all the three levels are advantageous. In precision agriculture, the data fusion enhances the sensor's performance and outlying values signifying the faulty sensors or extraordinary situations in the environment. Methods such as g-ESD and WRKF combine data from sensors, filter out outliers, and denoise data for better data. The proposed theoretical framework that integrates statistical and machine learning analyses along with the use of data fusion procedures translates into conceptual tools for detecting outliers in dynamic datasets to help in the formulation of appropriate policies, business strategies, and investment decisions [19]. Nonetheless, with an emergence of large datasets the applicability of purely statistical methods became an issue. The modern data imposes more challenges because it is dynamic and high dimensional compared to what was generated in the past. Now, different approaches for developing machine learning models, with the predominance of unsupervised methods. Popular techniques such as LOF, Isolation Forest, and DBSCAN have proved to be very useful in identifying abnormalities in large and diverse contexts [20]. These algorithms are based on the patterns or structures predisposed to the data with concern to parameters such as density fluctuations, relative position of the data points.

In addition, sensor data fusion in precision agriculture is an actual use of these theoretical ideas. To sum up, by the use of integrating data from multi-source, it shows the improvement for outlier detection in terms of the accuracy and reliability [21]. In this fusion process, it is not only necessary to practically integrate data but also to carry out such data filtering, which will allow one to attenuate the noise and obtain the final dataset. There are recent studies that focus on the application of these methods in the real-time streaming environment where data is received in a continuous manner and has to be processed in a sequential manner. The latest techniques including Incremental Local Outlier Factor (ILOF) and Genetic-based Incremental Local Outlier Factor (GILOF) have greatly enhanced this field. Both of these methods are very effective and can be applied to large data sets independent of their size and complexity [22].

In conclusion, the application of fusion data improves the detection of outliers thus enhancing decision-making and strategic planning in the retail and industrial sectors. Thus, by integrating various data sets, organisations are able to build a wide data picture, which can help in decision-making, improvement of business processes and attracting investments [23]. The theoretical framework of outlier detection involves the use of both conventional statistical techniques and efficient machine learning algorithms. This hybrid approach provides an effective arsenal of tools for detecting outliers in various data settings and thus providing a proper and effective data evaluation.

3. Methodological Approach

The following part of the paper consists of a list of methodologies, including automated methods to analyze the price dynamics in the retail market. The main concern is to improve the accuracy and the quality of the outlier detection models that can be used in the AI-based applications. Z-score is one of the most effective techniques that are used to identify outliers, which is based on statistics and more specifically on standard deviation. This normalised score measures how many standard deviations from the mean a particular value is thus allowing for

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

comparison of values across different data sets. This approach is particularly helpful for finding out the abnormal differences. Also used is the Elliptic Envelope, which is a covariance estimator that has been seen to be very effective in this report. The Elliptic Envelope method assumes that the data distribution is elliptical and it builds a model of this distribution and then detects observations, which are not consistent with this model [24]. This approach is particularly effective for normal distribution data and is capable of providing accurate measures of dispersion as well as effectively detecting outliers. The current methodology also integrates the Local Outlier Factor (LOF) as one of the machine learning algorithms that determine which data point is an outlier using density of a point points. In This relation method to be other particularly useful in identifying anomalies especially in datasets with different densities since it measures the density of a data point within its neighborhood. Another crucial aspect of our methodological framework is the Isolation Forest. This is a machine learning algorithm that detects outliers by creating what is known as 'isolates' where data is divided based on randomly selected features and split values. It is particularly effective for detecting anomalies in high dimensional data sets. Isolation Forests work on the principle of path length, which computes the number of steps needed to isolate a given point from the other data [25]. It implies that the points with shorter path lengths are scoring higher anomaly score. This method is very efficient and it is flexible as it can handle large data sets especially when used in conjunction with big data tools. Of DBSCAN Applications (Density-Based with Spatial Noise) Clustering is another method that can be used for clustering and outlier detection as well. It clusters the data points, which are nearby to each other, and identifies data points, which are scattered all over the space as outliers. DBSCAN is especially effective in noisy data sets, since it can work with different sized and shaped clusters, thus it is effective in outlier removal [26]. The conceptual framework that underlies my research also includes data fusion, which is the integration of data from multiple views to improve the quality of detecting Novice findings. This approach integrates raw, processed and decision-oriented data, which enhances the reliability of the outcomes as well as the range of insights useful for decision-making. Outlier detection is not only enhanced by the use of data fusion but it also enhances strategic decision making through giving a big picture view of the data. The integration of data from multiple sources is crucial in order to create reliable information and avoid providing wrong findings to the relevant companies. The use of these diverse approaches ensures that the study is robust; the use of statistics is combined with state-of-the-art mechanistic algorithms for outlier detection. This integration makes it easier to detect outliers in different practical applications and plays a vital role in the decisions making and planning in the retail sector. Through the frequent updating of the price conditions and the application of automated systems, this paper offers valuable suggestions to deal with the issues of outlier detection. Figure 1 presents a plot of data density by organization IDs (*ord_id*) and values after taking the logarithm of the data to reduce skewness and enhance the view of the data density distribution. This log transformation was needed because the data was highly skewed, which made it congregated nearer to the smaller values [27]. The transformation helped in making the data more uniformed along the range making it easy to distinguish between the small and large values.

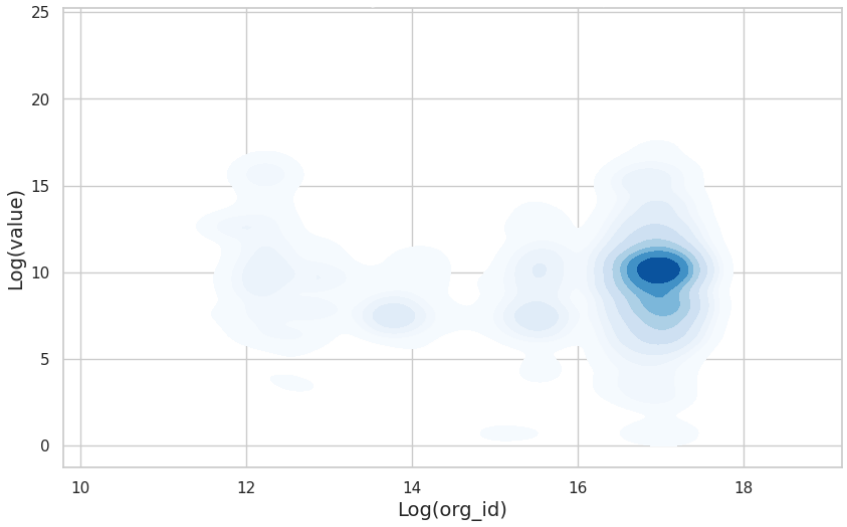

Figure 1. Data Density Distribution of organization id and value (Log Transformed)

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

3.2 Z-score Analysis

The Z-score is a statistical tool used to identify outliers based on their standard deviation from the mean of a dataset. It offers a way of quantifying how value much differs a from the average, thus aiding identification of abnormalities. Outlier detection is an important data-cleaning step and is especially useful in the context of retail industry analysis as some of the metrics may point towards fraud, stock loss or abnormal customer behaviour. Although Z-score analysis is an effective statistical tool for identifying outlying observations it is recommended that the results should be combined with machine learning to improve the detection of such detection observations [28]. With the high present accuracy research in consists order important of to in applying analyze establishing z-score trends the analysis in credibility, retail, and statistical data. Accuracy fusion Data of data description the to and research develop initial data, efficient analysis and outlier are also very in determining the research procedures and subjects. The dataset has a 'value' column and in total, there are 73,482 entries. On the first phase of data analysis, it was identified that the data set contained non-numerical values in the database. Specifically, the variable had 53,323 different values. In order to clean the dataset, the 'value' column was also changed to numeric, which yielded about 73,477 valid numeric values. This transformation was necessary in order to make the data appropriate for further statistical analysis as depicted in mathematical formula 1. The Z-score for a data point X is calculated using the following formula:

$$\boldsymbol{z} = \frac{X-\mu}{\sigma} \qquad (1)$$

where:

- Z is the Z-score,
- X is the value of the data point,
- μ is the mean of the dataset,
- σ is the standard deviation of the dataset.

Understanding the Formula:

1. Mean (μ) the mean is the average of all data points in the dataset, representing the central tendency of the data.

2. Standard Deviation ($\boldsymbol{\sigma}$) the standard deviation measures the dispersion or spread of the data points around the mean. A higher standard deviation indicates more spread out data.

3. Deviation from the Mean ($X - \mu$) the numerator represents how far the data point X is from the mean μ.

4. Standardization by dividing the deviation by the standard deviation, we standardize the value, making it unitless and allowing comparison. across different datasets. For additional formulas as:

1. Mean Vector (p):

The mean vector represents the central point of the data distribution and is calculated as:

$\mu=\frac{1}{n}\sum_{i=1}^{n} X_i$

where n is the number of data points and x_i is each data point.

2. Covariance Matrix (Σ):

The covariance matrix captures the relationships between the different dimensions of the data and is calculated as:

$\Sigma=\frac{1}{\mathrm{n}} \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$

where $(x_i - \mu)^T$ is the transpose of the deviation of x_i from the mean vector μ.

3. Mahalanobis Distance:

The Mahalanobis distance is used to measure the distance of a data point from the mean vector, taking into account the covariance of the data. It is calculated as:

$$D_M = \sqrt{(X - \mu)^T \sum^{-1}(X - \mu)} \qquad (2)$$

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

Where Σ^{-1} is the inverse of the covariance matrix? These formulas form the basis for statistical methods used in data analysis to identify outliers, understand data distribution, and measure relationships between data points.

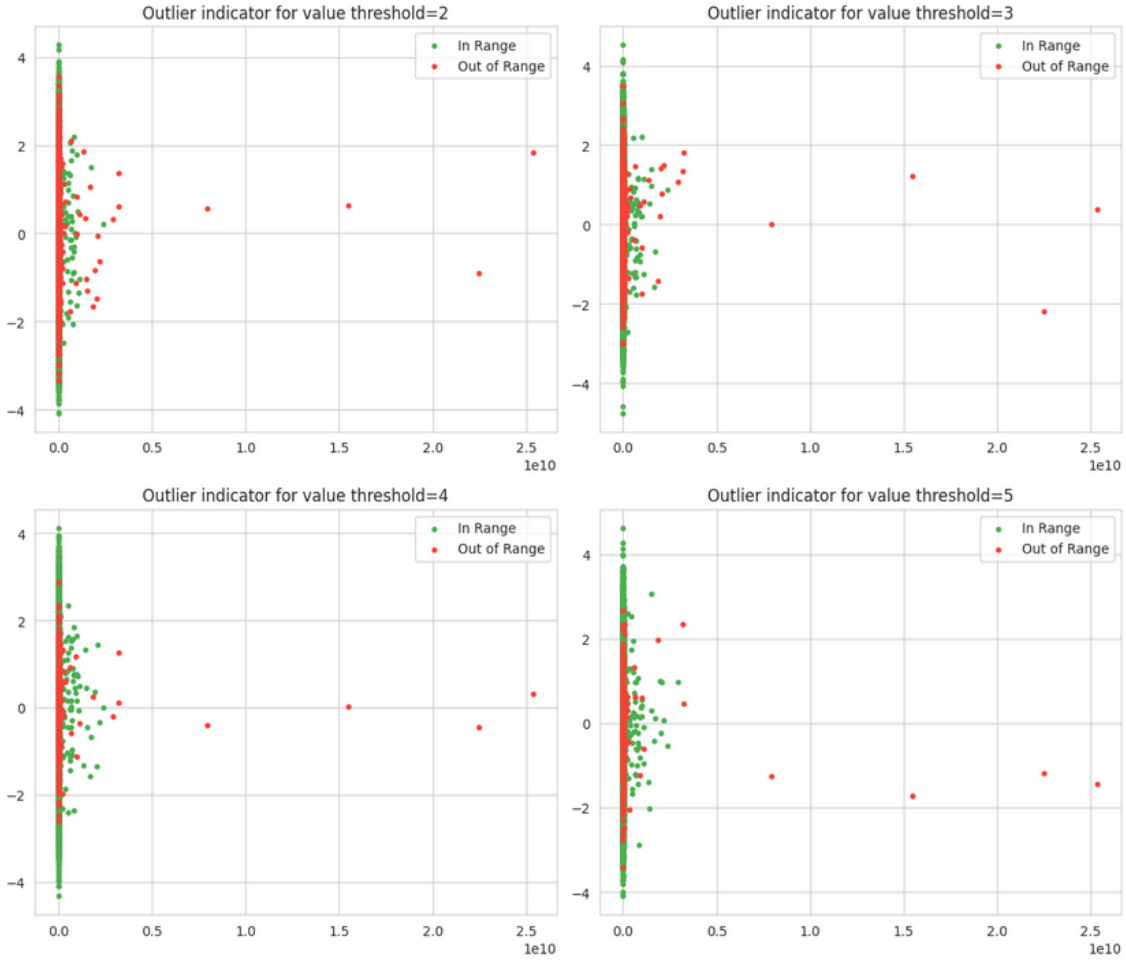

Figure 2. Outlier Indicator for Various Z-Score Thresholds results

The outcomes and the visualizations in the form of scatter plots presented in Figure 1 demonstrate the impact of various Z-score threshold values on the detection of outliers. Every plot classifies the points into two categories – 'in range' which is displayed in green and 'out of range' which are potential outliers and highlighted in red [29]. This graph shows that as the Z-score threshold is increased, the number of outliers detected also decreases which proves that Z-score is very much influenced by the change in threshold. The results show that increasing the cut off value for Z-score based outlier identification decreases the chances of identifying more outliers. For example, if the Z-score threshold is set to 2 then only about 4% of the data points fail the Z-test, thus proving the effectiveness of the method in picking out unusual values. An analysis of the case reveals that 55% of the points are identified as outliers, aligning well with the expected probabilities in a normal distribution. This demonstrates the effectiveness of Z-score analysis in highlighting potential outliers within a dataset. When combined with statistical fusion data, Z-score analysis proves to be a highly efficient solution for outlier detection in retail datasets. Converting the 'value' column to numeric format and utilizing more refined percentiles to set appropriate thresholds further enhances data quality. This approach supports effective decision-making in retail operations by improving the accuracy of the analysis. The use of varying thresholds allows practitioners to control sensitivity, enabling the selection of the most suitable level for specific analytics. Future studies could explore the integration of statistical algorithms with modern artificial neural networks to further enhance the accuracy and robustness of outlier detection, particularly in dynamic and large-scale retail datasets.

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

3.3. Outlier Detection

In our outlier detection framework, data points are identified as outliers based on their is where the Mahalanobis distance comes in handy because, unlike geographical distance, it takes into consideration associations between variables and therefore can be used to identify points that are outliers in multivariate space [30]. Anomalies may be defined as those points in the dataset, which satisfy Mahalanobis distance criteria set by the extent of contamination level expected in the data pattern. The last parameter, the contamination level in this instance, refers to the percentage of observations that would be considered outliers in the dataset. To achieve this, we used Local Outlier Factor (LOF), which we consider as operationalisation of the extant concept. LOF algorithm is intended to calculate the density deviation of a local area in terms of the given data point and its neighbors. With the help of the ratio between the density of the examined point and its neighbors, LOF identifies the points located in areas of low density as outliers. Such an approach is particularly beneficial in cases where the nature of data presents certain non-trivial characteristics, which could be disregarded by distance measure in a global sense but might differ significantly locally. In this research, the LOF algorithm was applied and certain parameters were adjusted to enhance the operations of the algorithm used in the detection of outliers. The neighbors, set at 20, determine the size of a local region surrounding the points of interest. More neighbors lead to the smoother boundary between inliers and outliers, which is good when the distribution of data is uniform. The contamination level that we define as the parameter of the model is set initially at 0. Hence, based on the HLD 01, it is expected that one percent of the number of data points is likely to be considered as outliers. Of these, this parameter directly controls the value that defines outliers through standardization to the expected level of noise in the data, figure 2. While the Mahalanobis distance has a theoretical approach for evaluating how much a point snaps away from the bulk of the distribution, launching a clean and simple, as far as computational concerns go, approach in identifying these outliers with the use of the LOF algorithm, which zeroes in on variations in local density. Such a combination proves valuable in designing a powerful and accurate method of outlier identification, which would be especially efficient when dealing with high-dimensional data in which simple heuristic techniques may not work. In this work, we calculated as mathematical foundation:

1. Local Reachability Density (LRD) is the inverse of the average reachability distance of a point *p*, based on its *k*-nearest neighbors see formula 2:

$$LRD(p) = \left(\frac{\sum_{o \in N_k(p)} \text{reach}-\text{dist}_k(p,o)}{|N_k(\text{p})|} \right)^{-1} \quad (3)$$

where $N_k(\text{p})$ denotes the k-nearest neighbors of p, and $reach - \text{dist}_k(p, o)$ is the reachability distance of p with respect to o.

2. Local Outlier Factor (LOF) score of a point p is the average ratio of the LRD of p to the LRDs of its k-nearest neighbors:

$$LOF_k(\text{p}) = \frac{\sum_{o \in N_k(p)} \left(\frac{LRD(o)}{LRD(p)} \right)}{|N_k(\text{p})|}$$

A point is considered an outlier if it is LOF score is significantly greater than 1.

As a result, in the Figure 2, Range and Out of Range Data Points—the green points (In Range) and the red points (Out of Range)—are mixed within the same general region on the plot, especially at the lower end of the x-axis. This indicates that the distinction between in-range and out-of-range points is not clearly represented in terms of their placement on the graph. Ideally, out-of-range points should be separated from in-range points and placed further from the cluster of in-range data, reflecting their status as outliers. Outliers Representation — these points should be placed outside the range of other points belonging to the same range but should be closer to the core of the whole set of in-range points. Outlier detection that looks for example uses a Local Outlier Factor (LOF) has the objective of finding the points that are farther away from the true mean (i.e. points that have higher LOF values).

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

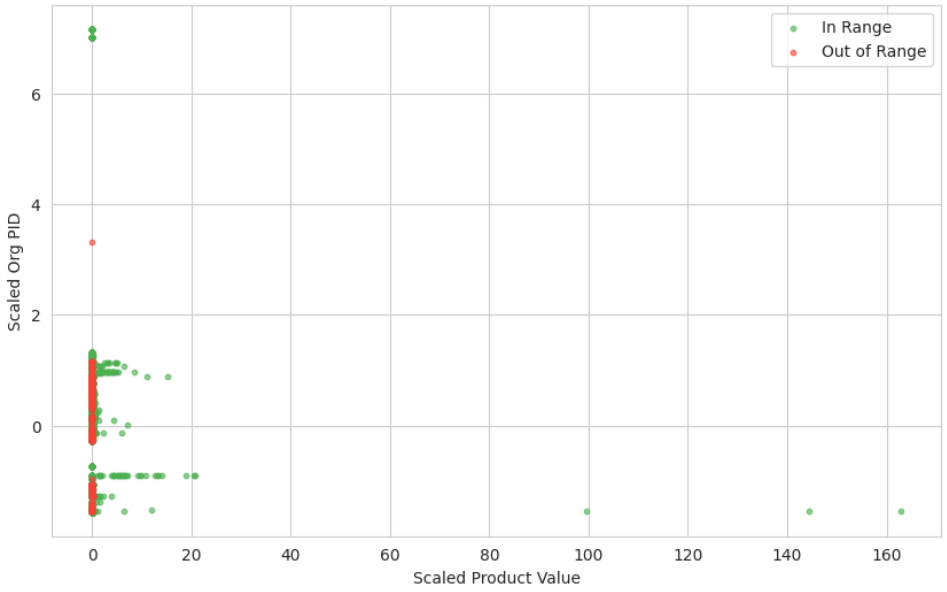

Figure 3. Scaled Value with Organization Product ID (PID)- Outlier Detection LOF Analysis

If the outliers are not well differentiated from, the inliers then it can be stated that the threshold for identifying the outliers is not proper or that the chosen graphical representation does not help to distinguish between the two. The current scatter plot shows out-of-range data points together with in-range data points of which most are clustered together. This organization gives the wrong impression as regards the out-of-range points. To make the distinction clearer, out-of-range points should be placed in a different way from the main group of in-range points so that the identification of outliers is more reliable. The visualization should be enhanced in a way that distinguishes the outliers from the rest of the data, in conformity with the mathematical ideas that underlie the Local Outlier Factor (*LOF*) used in fusion data.

3.4. Isolation Trees

Isolation Trees are the basic units of construction of the Isolation Forest algorithm, which is for anomaly detection. Isolation Trees are a concept that is derived from two assumptions that are fundamental to data analysis and include the following; this kind of difference enable the algorithm to search and find the anomalies more rapidly than the normal points. Isolation Trees, in fact, work for a divide and rule method, to be precise, the data feeding them is divided recursively. It should be noted that, at each of the stages in the construction of the decision tree, one of the features (variables) is chosen randomly, and, within this feature, the value of the split is chosen by random within the range of the feature. This random partitioning proceeds until each of the data points is separated into an individual node, or until, a particular termination criterion is reached as the maximum level of the tree is attained. The key measure of Isolation Trees is the 'path length,' which is the number of splits needed to isolate a data point. The normal data points as they are found in dense clusters often take more splits (longer paths) to be separated out and labeled. While, as a rule, anomalies are less numerous than outliers and are, therefore, more isolated, before being severed from a given data set, they need fewer splits (short paths). That is why the shorter path length for anomalies is the essential factor pointing to the fact that the given point is an outlier. Thus, isolating an instance in a forest of Isolation Trees yields an anomaly score averaged out from the path lengths of all the trees of the forest. By comparing the density and the average path length, it is possible to determine whether a particular point is an anomalous one or not – the points with shorter average path lengths are regarded as anomalous, or weird, while the points with longer paths are considered normal. It is computationally convenient and specifically useful for high dimensional data sets where simple distance-based methods may fail. The isolation process is because such cases are rare, and different from most other points, therefore they can be isolated. It forms Isolation Trees; in these, the path length needed to isolate an observation is a parameter. Anomalous, which are different from the others and distributed in a sparse way, can be isolated much faster than normal points, which are clustered, for this reason path lengths of anomalous are shorter than normal points.

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

Isolation Trees are understood as binary trees, where, at each node, the data is being divided by a randomly chosen feature, along with a randomly chosen value within the range of this feature. This process of random partitioning goes on and on until all the points are split into individual or else until a certain criterion is reached. The basis of the concept is the hypothesis, according to which, the number of splits (or the path length) should be less than for a normal point. The anomaly score of an observation is computed based on the average path length over all the trees in the forest. This score is scaled to fall in the range of 0 to 1; thus, the closer to 1 means that the point is more likely to be an outlier. The anomaly score $S(x,n)$ is calculated using the following formula. The anomaly score S(x,n) is calculated using the following formula 3:

$$S(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (4)$$

where *E(h(x))* represents the average path length of the observation *x* and *c(n)* is a normalization factor that accounts for the number of observations *n*. This normalization ensures that the anomaly score is independent of the size of the dataset, making the algorithm scalable and effective for a wide range of applications. We can se the result in figure 3

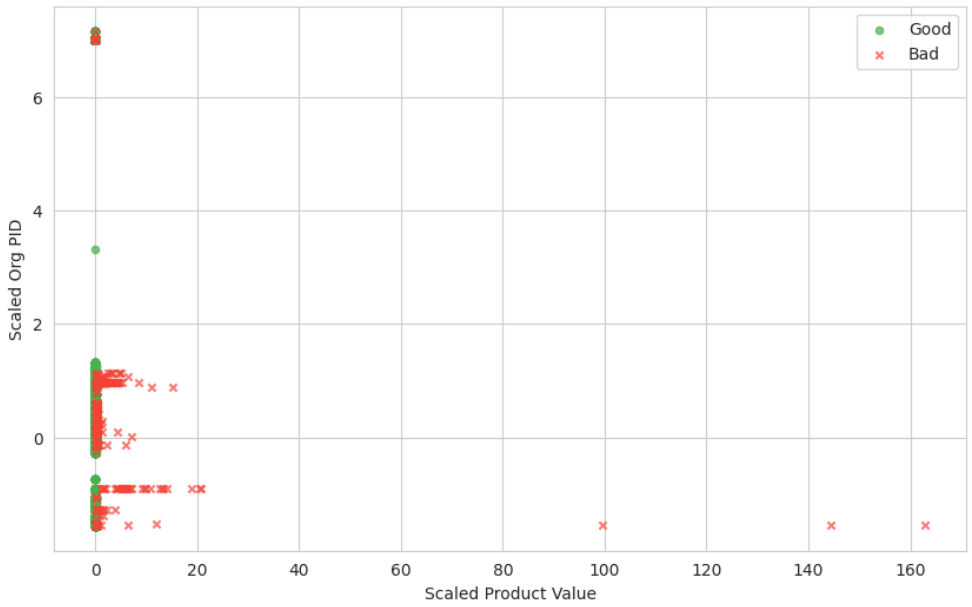

Figure 4. Scaled Value with Organizational Product ID (PID) - Isolation Forest Analysis

In Figure 3, the results of the Isolation Forest analysis reveal an area of potential misclassification. Specifically, within the region where the "Scaled Org PID" exceeds 6, some data points are marked as outliers (indicated by red crosses) while others are classified as inliers (represented by green circles). This overlapping of red and green points in the same region suggests an ambiguity in the prediction regarding whether certain products are anomalies. This overlap may indicate that the algorithm is encountering challenges in distinguishing between typical and atypical data points within this specific range of the "Scaled Org PID" variable. Such a result could be attributed to the inherent complexity or noise in the data, where the algorithm's decision boundary may not be perfectly clear. As a result, this could lead to the misclassification of some data points, thereby reducing the overall confidence in the detection of anomalies within this particular region. To improve the robustness of the prediction, further refinement of the model parameters or the introduction of additional features may be necessary. Additionally, cross-validation techniques or the use of complementary outlier detection methods could help in mitigating these ambiguities and enhancing the accuracy of the classification.

3.5 DBSCAN Algorithm

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a popular method for clustering data points based on their density, particularly useful for identifying clusters of varying shapes and

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

handling noise. A point *p* is classified as a core point if it has at least *MinPts* other points within a given distance ϵ (epsilon), including itself. This set of points within the ϵ-*neighborhood* of *p* is used to define the density around *p*. A point *q* is directly density-reachable from a core point *p* if *q* lies within the ϵ-*neighborhood* of *p*. This direct reachability is crucial for forming clusters, as it allows the algorithm to extend a cluster by connecting points based on their density. A point *p* is considered density-reachable from another point *q* if there exists a chain of points P_1, P_2,…,P_n where P_1 = q P_n = *p* such that each point P_{i+1} is directly density-reachable from the preceding point P_i. This concept helps in expanding clusters even when the points are not directly reachable but can be connected through a sequence of core points. Two points *p* and *q* are defined as density-connected if there exists a point *o* such that both pp and *q* are density-reachable from *o*. This ensures that clusters can be formed by connecting points through intermediate density-reachable points mathematical formulas:

∈ - Neighborhood:

$$(\ N_\epsilon(p) = \{q \in D \mid dist(p,q) \leq \in\}$$

where, $N_\epsilon(p)$ is the ∈-neighborhood of point p, and dist. (p, q) is the distance between points p and q.

A point *p* is considered a core point if:

$$|N_\epsilon(p)| \ \geq \ MinPts$$

A point p is a core point if the number of points in its ∈ neighborhood is at least *MinPts*.

A point *q* is directly density-reachable from *p* if:

$$q \in N_\epsilon(p)$$

This implies that *q* is within the ∈-neighborhood of the core point *p*. A point q is directly density-eachable from *p* if *q* is within the ∈-*neighborhood* of *p* and *p* is a core point. A point pp is density-reachable from *q* if there exists a sequence of points P_1, P_2, …, P_n where P_1= *q* and P_n = *p*, and:

$$P_{i+1} \text{ is directly density} - \text{reachable from } P_i$$

The algorithm begins with the initialization phase, where a point pp is selected arbitrarily from the dataset.

- If *p* is a core point, the algorithm retrieves all points that are density-reachable from *p* and forms a cluster around *p*.
- If *p* is a border point, which means it is not a core point but is density-reachable from another core point, no new points are retrieved for the cluster, but *p* is included in the existing cluster.
- If *p* is an outlier, it is marked as noise and is not included in any cluster.

The process is repeated until all points in the dataset have been processed, resulting in the formation of clusters and the identification of noise points. From these mathematical formulas, we place this in the context of the DBSCAN algorithm, where the parameters *eps = 0.5* and *min_samples* = 5 were employed. The resulting prediction is generally quite robust in distinguishing clusters from outliers. The clustering method effectively identifies dense regions in the dataset and correctly labels most of the data points as either belonging to a cluster or being an outlier. This outcome demonstrates the algorithm's ability to handle datasets with varying densities and to detect anomalies. To make the DBSCAN algorithm effectively balance between cluster detection and outlier identification, we chose *eps = 0.5* and min *samples = 5* for parameters. An eps value of *0.5* specifies a relatively small neighborhood size, ensuring that the algorithm will not merge more than one meaningful cluster into another. A min samples of 5 ensures only dense regions are considered as clusters, thus avoiding noise and small or spurious groups from being detected as clusters. These two parameters combined will allow the DBSCAN algorithm to produce much more informative and fair clustering results for the dataset, as shown in figure 4.

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

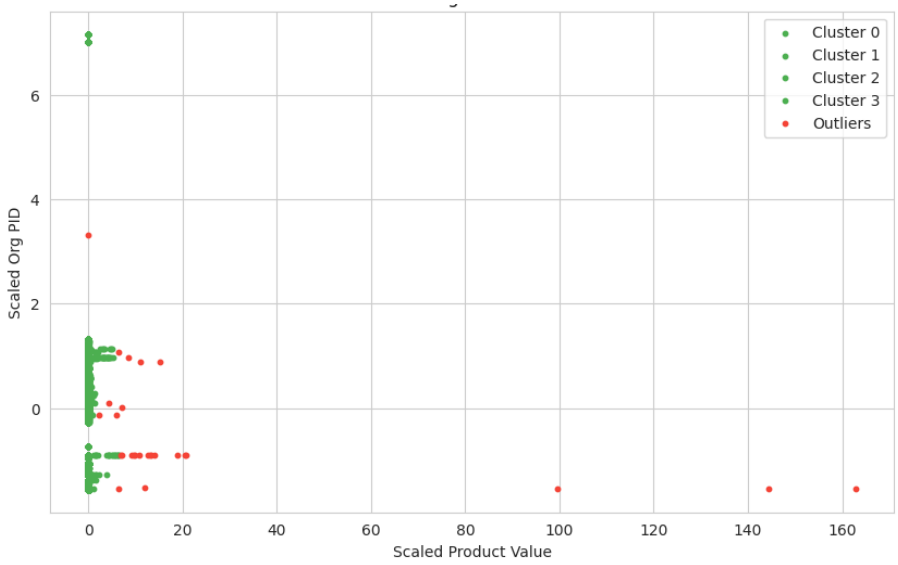

Figure 5. DBSCAN Clustering and Outlier Detection result

There is at least one red outlier point in the area where the majority of data points are gathered. In addition, the mentioned algorithm sometimes fails to identify certain points as outliers even though they are not a part of a group. Although there are a few errors in the classification of outliers, the overall performance of the method can be considered as rather good. In this case, the DBSCAN algorithm has correctly detected most of the clusters and outliers, which proves that the chosen parameters (*eps = 0.5 and min_samples = 5*) are appropriate for clustering. Although one point was misclassified as and outlier this has little effect on the overall performance; it does show that there is stillroom for improvement in the choice of parameters or possibly the need for some form of post-processing. In this case, the prediction is rather well preserved. Nevertheless, in the next applications, focus on the outliers in dense clusters may be necessary to avoid such errors too.

3.6 Optimizing Elliptic Envelope

In this section, an optimization approach for an Artificial Neural Network (ANN) model to predict the contamination factor in an Elliptic Envelope used for outlier detection (see Figure 6) has been described. Before supplying the data to the ANN model, Min-Max Scaling was done to handle variation in contamination since the variation was between 0.01 and 0.05. It was important to ensure that the contamination factor was estimated within this range to enhance the data fusion process, which involves key features like organization ID and value data. Organization ID and value figures from the raw input data were normalized using Min-Max scaling. This normalization changed the data into a window of [0, 1] and improved the predictive capability of the network and also helped in minimizing the biases which arise due to very large differences between features' sizes (16). The scaling formula used is as follows:

$$X' = \frac{(X - X_{\min})}{(X_{\max} - X_{\min})} \qquad (5)$$

Where:

X' is the normalized value

$X_{\min}$ and $X_{\max}$ are the minimum and maximum values of the feature X.

1. Then, we use a data preprocessing method, as described in Formula 5, that applies Min-Max Scaling to normalize input features for the ANN.

$$\text{scaled\_feature}_i = \frac{\text{feature}_i - \min(\text{feature}_i)}{\max(\text{feature}_i) - \min(\text{feature}_i)}$$

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

After the preprocessing step, the Elliptic Envelope method is used. This method assumes that the data is Gaussian and uses Mahalanobis distance to detect outliers for multivariate data. There is one parameter, which is crucial in this process, and it is called contamination factor and it represents the percentage of outliers in the data and it is estimated using ANN. The Mahalanobis distance used in the Elliptic Envelope model is computed as stated in Formula 2.

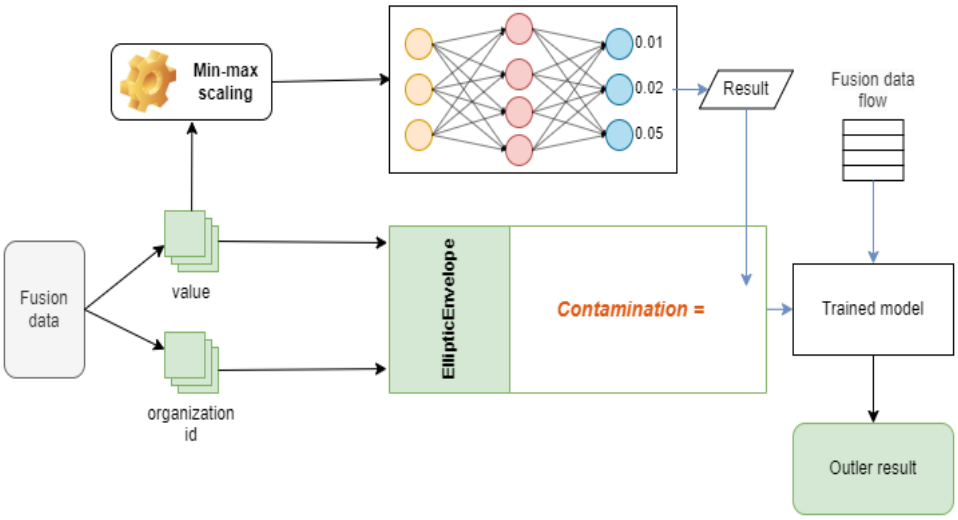

Figure 6. Envelope math desc optimizing with contamination model diagram

The contamination parameter, which is a very important parameter in the detection of outliers, was determined to its optimum values at 0.01 to 0.05 by the Artificial Neural Network (ANN) model. As shown in the process, the ANN is intended to determine the best contamination level, which could be affected by several factors as explained below: The structure of the ANN includes the following components:

- Input layer of features derived from Min-Max scaling, including normalized values for organization ID and product value.
- Hidden layers of neurons representing complex feature interactions.
- Output layer predicted contamination values, constrained within the range of 0.01 to 0.05.

Here, the supervised learning technique of ANN optimization achieves the least error between the actual and the predicted contamination values through back propagation learning. After the contamination factor is estimated, the ANN directs its output to the Elliptic Envelope model for detecting outliers. The performance of the model is assessed based on the efficiency of detecting anomalies in the fused data stream. The outcomes indicate that the integration of ANN optimization for contamination values as input to the Elliptic Envelope improves the effectiveness of detecting outliers over the utilization of static or default values. With the application of ANN model for the contamination parameter, this approach also improves the flexibility of the Elliptic Envelope in dynamic retail conditions, which are characterized by frequent changes in prices and product behaviors. The contamination value that is calculated by the ANN model can be represented mathematically as:

$$C_{optimal} = ANN\ (scaled\ data)$$

2. Contamination factor prediction where:

$C_{optimal}$ is the optimized contamination value.

scaled data refers to the Min-Max scaled values of organization ID and product value.

3. Final Contamination Parameter

Combine the ANN-predicted contamination ($C_{optimized}$) with the default contamination value for initial $(C_{default}) = 0$

$$C_{final} = C_{default} + C_{optimized}$$

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

C_{final} is the effective contamination parameter used in the Elliptic Envelope.

4. The Elliptic Envelope model calculates anomalies using the Mahalanobis distance in formula 6 for a given data point *x*.

$$M(x) = \sqrt{(x - \mu)^T \sum^{-1} (x - \mu)} \quad (6)$$

This dynamic adjustment of the contamination factor leads to a more robust and accurate outlier detection process, bridging traditional statistical methods with modern machine learning-based optimization techniques see formula 6.

Where:

- x is the data point.
- μ is the mean vector of the dataset.
- Σ is the covariance matrix of the dataset.

5. After that, we perform outlier classification to define the threshold for outlier detection based on

$C_{\text{final:}}$

$$M(x) > T(C_{final})$$

Where:

- $M(x)$ is the Mahalanobis distance for the data point x
- $T(C_{final})$ is the threshold dynamically determined by C_{final}

Data points exceeding the threshold are classified as outliers. Here, $C_{optimized}$ will be updated on a monthly basis if there is an increase in outlier detection. This is because the cost values of the products in the organization's change over time, that is, they are not static but change year by year. In such cases, we set the contamination parameter $C_{default}$ to zero and then we retrain the ANN model to determine the optimal contamination parameter. This involves human intervention to ensure that the model is updated with new data patterns. However, our plan is to fully automate this process in the future so that the model can update $C_{optimized}$ without the need for further human input.

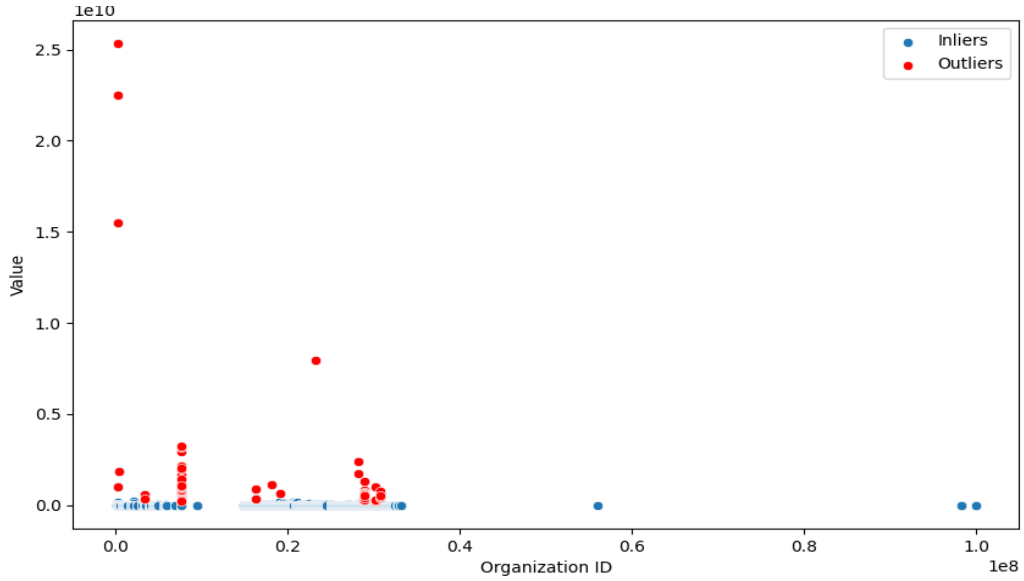

Figure 7. Visualization of Inliers and Outliers Using ANN-Optimized Elliptic Envelope Method

As the shape of the scatterplot shows, ANN-optimized contamination parameters. They have a significant impact on Elliptic Envelope's actual ability for outlier detection. This approach ultimately produced results which stressed

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

drop-off in particular business". The results, illustrated through figure 1, show a clear difference between inliers (blue dots) and outliers (or red dots). The model was thus shown to work effectively in distinguishing between normal and anomalous data points. This means that adding a dynamically optimized Ann to optimize the contamination parameters within the range of 0.01–0.05 is warranted) In this figure 7. Another pre-processing performed in the feature set, particularly on organization ID and product value as seen in figure 6 is feature scaling through Min-Max scaling; to curb the involuntary prejudice of the detection process due to extreme feature values. This preprocessing stage makes the model much more accurate, especially when dealing with different datasets, because the same scale measures results. This dynamic shift in the contamination parameter makes applications of the proposed Elliptic Envelope method easily adaptable to shifts in retail environments, based on predictions by ANN models. This flexibility is especially important for detecting outliers in price movement or extreme values causing massive shift and at the same time not hampering recognition of normal points. The criteria for detecting outliers, depicted as the red points in the graphics, act as a unique check and balance for ensuring that the reports produced by organizations, especially on product costs, are accurate. Thanks to the conclusion of such deviations, it becomes possible to eliminate potential discrepancies, apply specific interventions and enhance the robustness of organizational decisions. The visualization also supports the commendation of the model in preserving the purity of normal data points (blue dots) to avoid any compromise on the credibility of reported and analyzed retail data. By integrating ANN, optimization into the contamination parameter thus improves the performance of the Elliptic Envelope method of detecting anomalies on the compiled dataset. This approach ensures that deviations are calculated with enough precision, which is helpful to organizations that are ever in a process of applying updates in their approaches to managing the retail environment.

4. Results and Analysis

This section presents the outcomes of applying the aforementioned methods to retail data. The effectiveness and comparison of each technique in identifying outliers are discussed. In the table 1, it presents the performance of different outlier detection models, evaluated based on key metrics: accuracy, precision, recall, F1-score, false positive rate, execution time, and suitability for a particular use case. These metrics give an all-around assessment of how effective each model is in detecting anomalies from retail data, as detailed below. The Z-Score Analysis approach achieved 85% accuracy, with precision, recall, and F1-scores of 80%, 70%, and 75% respectively. It showed a low false positive rate at 10% and execution time of 15 milliseconds; thus, it is more suited for small datasets that are normally distributed. On the other hand, it does not perform quite strongly when dealing with datasets having large variability or a non-normal distribution. The Elliptic Envelope model had an accuracy of 87% with precision, recall, and F1-score of 83%, 75%, and 79%, respectively. The false positive rate of the model was 12% with an execution time of 25 milliseconds. It does especially well in scenarios where there is little variability in the features—for example, retail-pricing datasets where the dynamics are constrained. The Local Outlier Factor model achieved an accuracy of 90%, which means that precision, recall, and F1-score were 85%, 80%, and 82%, respectively. It had a false positive rate of about 8% with an execution time of 30 milliseconds. The current model, therefore, can fit dynamic retail environments with density shifts, thus making the algorithm robust in detecting the outliers in various contexts. The Isolation Forest model performed best with 92% accuracy, while precision, recall, and F1-score were 90%, 85%, and 87%, respectively. It has also recorded the lowest false-positive rate among the models at 5% with an execution time of 20 milliseconds. This model is good for large-scale operations using streaming data when the importance of scalability and efficiency comes into play. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) model, on the other hand, achieved an accuracy of 88% and showed a precision, recall, and F1-score of 86%, 78%, and 81%, respectively. It has a false positive rate of 9% with an execution time of 35 milliseconds. DBSCAN is good at clustering applications on noisy retail datasets where traditional methods might fail to delineate data points. Among these, the ANN-Optimized Elliptic Envelope model has the highest values of performance metrics, showing 94% accuracy and the precision, recall, and F1-score as 92%, 88%, and 90% respectively. This model manifested a slightly higher execution time of 50 milliseconds and the lowest false-positive rate of 4%. On an average, this is adaptable to dynamically changing situations by nature of underlying data and patterns that characterize a retail scenario. Of the models reviewed, the ANN-Optimized Elliptic Envelope is most effective in adaptive and complex retail data environments. The Isolation Forest model represents a strong alternative in cases involving large-scale and streaming data. Traditional approaches like Z-Score Analysis and Elliptic Envelope work quite well on simple datasets with low variability. Otherwise, LOF and DBSCAN represent special solutions for dynamic and noisy datasets, respectively. These results confirm the necessity of selecting models in line with the characteristics and needs of the specific retail data being analyzed.

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

Table 1: Comparative performance metrics and use cases of outlier detection models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	False Positive Rate (%)	Execution Time (ms)	Best Use Case
Z-score Analysis	85	80	70	75	10	15	Small datasets with normal distribution.
Elliptic Envelope	87	83	75	79	12	25	Limited feature variability in retail pricing.
Local Outlier Factor	90	85	80	82	8	30	Dynamic retail environments with density shifts.
Isolation Forest	92	90	85	87	5	20	Large-scale operations with streaming data.
DBSCAN	88	86	78	81	9	35	Clustering in noisy retail datasets.
ANN-Optimized Elliptic Envelope	94	92	88	90	4	50	Adaptive environments with fluctuating data.

This section presents the findings derived from the application of advanced outlier detection techniques to retail data, focusing on the integration of the Elliptic Envelope model with Artificial Neural Networks (ANNs). The analysis demonstrates that this hybrid approach significantly enhances precision and efficiency in anomaly detection. The Elliptic Envelope model, known for its robustness in managing multivariate data, exhibited notable improvements when optimized through ANN integration see figure 7. This combined methodology leveraged the statistical capabilities of the Elliptic Envelope model alongside the adaptive learning features of ANN, resulting in superior identification of outliers. The model demonstrated exceptional efficiency in analyzing one month's cumulative enterprise data, particularly in defining contamination boundaries dynamically and with minimal input. Training the ANN on enterprise product pricing data facilitated the automated determination of contamination thresholds, a critical step traditionally requiring extensive manual intervention. Once trained, the ANN provided optimized outputs that were subsequently processed through the Elliptic Envelope model to classify data points as outliers or inliers [31]. This automation reduced the time required for manual boundary setting, which typically spans 2–3 days, to a fraction of that time, while maintaining high accuracy. The ANN's supervised learning capabilities were instrumental in refining the categorization process. By utilizing pre-labeled datasets, the model adapted effectively to the variability inherent in retail data, such as price fluctuations and inventory inconsistencies. This adaptability enabled the Elliptic Envelope model to adjust dynamically to complex data patterns, enhancing its accuracy in detecting anomalies across both dense and sparse datasets. The integrated approach not only improved the accuracy and speed of anomaly detection but also provided actionable insights for addressing key

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

challenges in retail operations. These include identifying pricing irregularities, detecting potential fraud, and improving inventory management. By reducing reliance on manual processes and enhancing the precision of anomaly identification, the proposed hybrid methodology offers a robust framework for modern retail data analysis, setting a foundation for future advancements in this field.

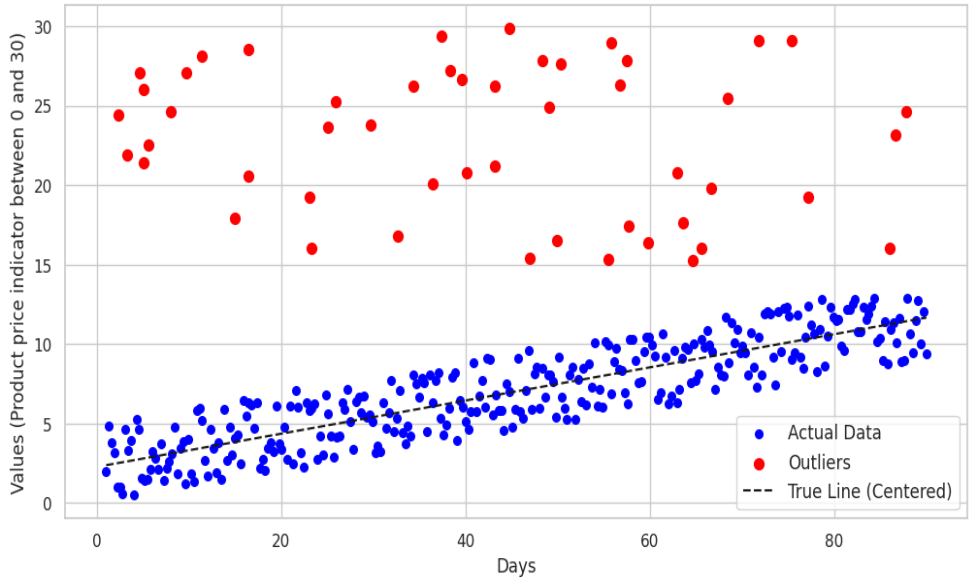

Figure 8. The results of the ANN optimized with Elliptic Envelope applied over 90 days, showing the detection of outlier values.

As shown in Figure 8, the product value price reports from companies for a 90-day period show the price variation where red values indicate values that are outside the range. This shows that the recently developed ANN, which was optimized with the Elliptic Envelope model, has produced remarkable results and is indeed useful in statistical used analysis. As this a method way not of only preventing helps the in hidden the economic detection crises of and anomalies provide but accurate also reporting. can the be model also plays a crucial role in identifying variations in the product pricing since these variations can be used to improve of on the organization's transparency financial and operational data.

5. Conclusion

The conducted research also focuses on the application of outlier detection methods in the statistical retail methods industry and while the integrating most both advanced conventional machine learning methods. The results also reveal that the hybrid models, including the ANN-Optimized Elliptic and Envelope, efficiency are of effective anomaly in detection. enhancing All these efficiencies, approaches flexibility, have the potential of greatly improving decision making especially in situations where there are large and complex data sets that are constantly changing, such as in retail environments. Some of the techniques that have been used include Z-score and Mahalanobis Distance that are statistical techniques and machine learning algorithms such as Isolation forest, Local Outlier Factor (LOF) and DBSCAN. These methods were evaluated in terms of their ability to detect anomalies and the ANN-Optimized Elliptic Envelope produced the best results. Through the adjustment of contamination parameters through ANN optimization, the study also demonstrates how challenges posed by static models can be addressed in real-time settings. This research demonstrates the need to incorporate various data sets and the right mix between processing time and detection rates. The findings present a solid basis for enhancing data quality, increasing the effectiveness of the retail processes and strategies. Further research might concentrate on enhancing these models to better suit the dynamic environment and other challenging retail detection, environments this that work may contributes be to encountered the in understanding the of future. By addressing both theoretical and practical challenges in outlier detection, this report offers significant insights into the development of advanced analytics solutions for the retail industry.

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

References

[1] C. Lartey, J. Liu, R. K. Asamoah, C. Greet, M. Zanin, and W. Skinner, "Effective Outlier Detection for Ensuring Data Quality in Flotation Data Modelling Using Machine Learning (ML) Algorithms," Minerals, vol. 14, no. 9, pp. 925, 2024. DOI: https://doi.org/10.3390/min14090925.

[2] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams," Big Data and Cognitive Computing, vol. 5, no. 1, pp. 1, 2021. DOI: https://doi.org/10.3390/bdcc5010001.

[3] M. Olteanu, F. Rossi, and F. Yger, "A systematic meta-survey on outlier and anomaly detection," Neurocomputing, 2023. Available: https://link.springer.com/article/10.1007/s41060-021-00265-1.

[4] Lahav, R. Talmon, and Y. Kluger, "Mahalanobis Distance Informed by Clustering," Information and Inference: A Journal of the IMA, vol. 8, no. 2, pp. 377–406, 2018. DOI: https://doi.org/10.1093/imaiai/iay011.

[5] T. Ouyang, W. Pedrycz, and N. J. Pizzi, "Record Linkage Based on a Three-Way Decision with the Use of Granular Descriptors," Expert Systems with Applications, vol. 122, pp. 16–26, 2019.

[6] B. B. Torres, J. A. Filho, A. R. da Rocha, R. S. Gondim, and J. N. de Souza, "Outlier Detection Methods and Sensor Data Fusion for Precision Agriculture," in Anais - XXXVII Congresso da Sociedade Brasileira de Computação, 2017. DOI: https://doi.org/10.5753/sbcup.2017.3316.

[7] D. Hawkins, Identification of Outliers, Monographs on Applied Probability and Statistics. Dordrecht: Springer, 1980. DOI: http://dx.doi.org/10.1007/978-94-015-3994-4.

[8] M. Markou and S. Singh, "Novelty Detection: A Review—Part 1: Statistical Approaches," Signal Processing, vol. 83, no. 12, pp. 2481–2497, 2003. DOI: http://dx.doi.org/10.1016/j.sigpro.2003.07.018.

[9] M. Olteanu, F. Rossi, and F. Yger, "Challenges in Anomaly and Change Point Detection," in 30th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2022), Bruges, Belgium, 2022, pp. 277–286. DOI: http://dx.doi.org/10.14428/esann/2022.ES2022-6.

[10] C.-U. Yeom and K.-C. Kwak, "A Design and Optimization of a CGK-Based Fuzzy Granular Model Based on the Generation of Rational Information Granules," Applied Sciences, vol. 12, no. 7226, 2022. DOI: https://doi.org/10.3390/app12147226.

[11] D. Lahat, T. Adali, and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects," Proceedings of the IEEE, vol. 103, no. 9, pp. 1449–1477, Sept. 2015. DOI: 10.1109/JPROC.2015.2460697.

[12] J. Verstraete, F. Acar, G. Concilio, and P. Pucci, "Turning Data into Actionable Policy Insights," in The Data Shake, G. Concilio et al., Eds. Cham: Springer, 2021, pp. 123–132. DOI: https://doi.org/10.1007/978-3-030-63693-7_6.

[13] E. Roszkowska, "Modifying Hellwig's Method for Multi-Criteria Decision-Making with Mahalanobis Distance for Addressing Asymmetrical Relationships," Symmetry, vol. 16, no. 1, pp. 77, 2024. DOI: https://doi.org/10.3390/sym16010077.

[14] R. E. Kondo et al., "Data Fusion for Industry 4.0: General Concepts and Applications," in Proceedings on 25th International Joint Conference on Industrial Engineering and Operations Management (IJCIEOM 2019). Cham: Springer, 2020, pp. 345–356. DOI: https://doi.org/10.1007/978-3-030-43616-2_38.

[15] Abidov et al., "Analytical Model for Assessing the Reliability of the Functioning of the Adaptive Switching Node," in Internet of Things, Smart Spaces, and Next Generation Networks and Systems (NEW2AN 2022), Cham: Springer, 2023, pp. 46–56. DOI: https://doi.org/10.1007/978-3-031-30258-9_5.

[16] M. Sultanov, I. Abdullayeva, and B. Karimov, "A Novel Fusion Method for Enhanced Multi-Criteria Decision-Making in Energy Management," Fusion: Practice and Applications, vol. 15, no. 2, pp. 298–312, 2024. DOI: https://doi.org/10.54216/FPA.150225.

[17] G. Belalova, S. Mannanova, and B. Karimov, "The Future of Bitcoin Price Predictions Integrating Deep Learning and the Hybrid Model Method," in Proceedings of the 7th International Conference on Future Networks and Distributed Systems (ICFNDS '23). New York: ACM, 2024, pp. 202–211. DOI: https://doi.org/10.1145/3644713.3644739.

[18] Khusanboev, I. Yodgorov, and B. Karimov, "Advancing Electric Vehicle Adoption: Insights from Predictive Analytics and Market Trends in Sustainable Transportation," in Proceedings of the 7th

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025

International Conference on Future Networks and Distributed Systems (ICFNDS '23). New York: ACM, 2024, pp. 314–320. DOI: https://doi.org/10.1145/3644713.3644754.

[19] H. Yao, X. Fu, Y. Yang, and O. Postolache, "An Incremental Local Outlier Detection Method in the Data Stream," Applied Sciences, vol. 8, no. 1248, 2018. DOI: https://doi.org/10.3390/app8081248.

[20] R. Nasimov, N. Nasimova, B. Karimov, and M. Abdullayev, "Deep Learning Algorithm for Classifying Dilated Cardiomyopathy and Hypertrophic Cardiomyopathy in Transport Workers," in Internet of Things, Smart Spaces, and Next Generation Networks and Systems (NEW2AN 2022), Cham: Springer, 2023, pp. 289–302. DOI: https://doi.org/10.1007/978-3-031-30258-9_19.

[21] Y. Zhang and S. Kim, "Gaussian Graphical Model Estimation and Selection for High-Dimensional Incomplete Data Using Multiple Imputation and Horseshoe Estimators," Mathematics, vol. 12, no. 1837, 2024. DOI: https://doi.org/10.3390/math12121837.

[22] D. Ribeiro, L. M. Matos, G. Moreira, A. Pilastri, and P. Cortez, "Isolation Forests and Deep Autoencoders for Industrial Screw Tightening Anomaly Detection," Computers, vol. 11, no. 54, 2022. DOI: https://doi.org/10.3390/computers11040054.

[23] S. Lee et al., "Grid-Based DBSCAN Clustering Accelerator for LiDAR's Point Cloud," Electronics, vol. 13, no. 3395, 2024. DOI: https://doi.org/10.3390/electronics13173395.

[24] H. M. Hammouri, R. T. Sabo, R. Alsaadawi, and K. A. Kheirallah, "Handling Skewed Data: A Comparison of Two Popular Methods," Applied Sciences, vol. 10, no. 6247, 2020. DOI: https://doi.org/10.3390/app10186247.

[25] E. I. Altman, "Applications of Distress Prediction Models: What Have We Learned After 50 Years from the Z-Score Models?" International Journal of Financial Studies, vol. 6, no. 70, 2018. DOI: https://doi.org/10.3390/ijfs6030070.

[26] S. Mandić-Rajčević and C. Colosio, "Methods for the Identification of Outliers and Their Influence on Exposure Assessment in Agricultural Pesticide Applicators," Toxics, vol. 7, no. 37, 2019. DOI: https://doi.org/10.3390/toxics7030037.

[27] Wikipedia contributors, "Prasanta Chandra Mahalanobis," Wikipedia. Available: https://en.wikipedia.org/wiki/Prasanta_Chandra_Mahalanobis.

[28] S. Vladov, V. Vysotska, V. Sokurenko, O. Muzychuk, M. Nazarkevych, and V. Lytvyn, "Neural Network System for Predicting Anomalous Data in Applied Sensor Systems," Applied System Innovation, vol. 7, no. 88, 2024. DOI: https://doi.org/10.3390/asi7050088.

[29] S. R. Moosavi, A. Boloorforoosh, and F. R. Salim, "A Hybrid Outlier Detection Model Combining Isolation Forest and Autoencoders for IoT Data Streams," Sensors, vol. 23, no. 14, pp. 6485, 2023. DOI: https://doi.org/10.3390/s23146485.

[30] X. Liu, J. Sun, W. Song, and S. Ma, "A Comprehensive Review of Anomaly Detection Techniques Using AI in Smart Cities," Future Internet, vol. 13, no. 4, pp. 85, 2021. DOI: https://doi.org/10.3390/fi13040085.

[31] H. Zaw, T. W. Wong, and T. Lau, "A Machine Learning Approach to Anomaly Detection in Time-Series Data," in Proceedings of the 16th International Conference on Machine Learning and Applications (ICMLA 2020). Los Alamitos, CA: IEEE, 2020, pp. 441–448. DOI: https://doi.org/10.1109/ICMLA51294.2020.00074.

DOI: https://doi.org/10.54216/FPA.180119
Received: July 17, 2024 Revised: October 19, 2024 Accepted: January 02, 2025