International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.8 March - April – 2024
DOI:10.46647/ijetms.2024.v08i02.020 ISSN: 2581-4621

Machine Learning Technique for Anomaly Detection

Victor Pitchaimuthu[1], Avinash Holla[2]
[1]*consultant, victor.p@hcl.com, HCLTech Ltd, Bengaluru*
[2]*Technical Architect, avinash.ks@hcl.com, HCLTech Ltd, Bengaluru*

Abstract
Anomaly detection is an essential component of storage monitoring systems, allowing irregular patterns, deviations, and outliers to be identified. The objective of the project is to design and develop an iannovative storage monitoring system empowered by machine learning techniques to enable real-time anomaly detection, proactive management, and efficient utilization of storage infrastructure resources. This tool aims to continuously monitor diverse storage metrics, identify abnormal patterns or behaviors, and provide actionable insights for administrators to maintain optimal storage performance, minimize downtime and enhance data integrity across various storage systems.
The key focus is leveraging advanced analytics to predict, detect, and address potential anomalies, ensuring the reliability, efficiency, and security of storage operations within dynamic and data-intensive environments.
Keywords: Anomaly detection, machine learning, storage infrastructure, proactive management, data-intensive environment.

Introduction

An anomaly is defined as a significant departure from what is expected within a dataset, which is called a "normal" pattern. The detection of anomalies in storage monitoring may indicate potential issues, abnormalities, or security threats.
Key Characteristics:
1. Dynamic Environments: Anomaly detection is crucial in dynamic storage environments where operational conditions, user behaviors, and data patterns can evolve over time.
2. Data-Intensive Nature: In storage systems dealing with vast amounts of data, anomalies may signify potential inefficiencies, errors, or security breaches.
3. Real-Time Responsiveness: The ability to detect anomalies in real-time ensures timely responses to emerging issues, minimizing the risk of downtime and optimizing storage performance.

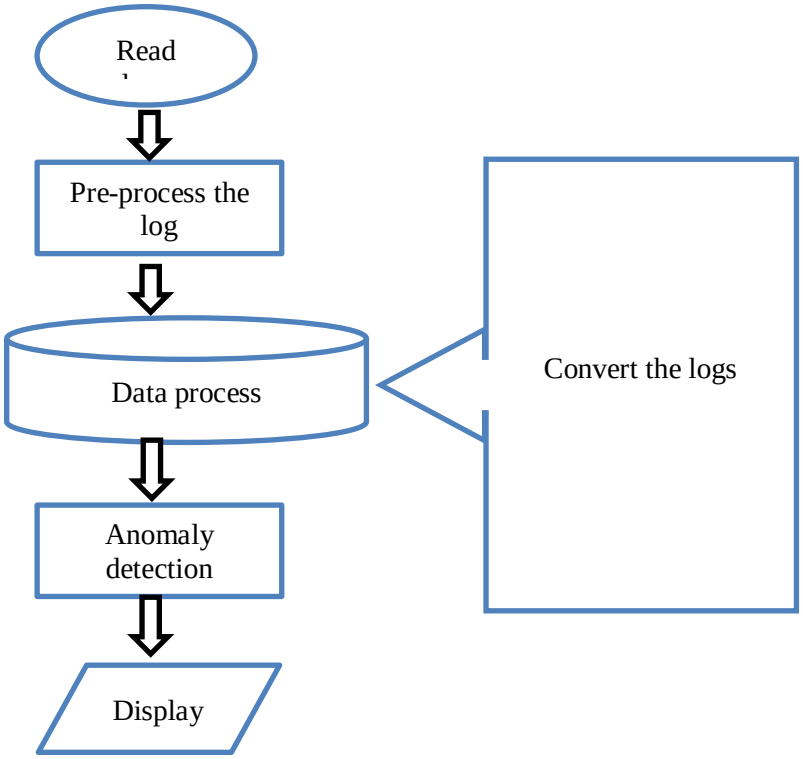
Model

Process flow:
The business process flow for the "Intelligent Storage Monitoring Tool" project involves various stages from data collection to user interaction.

Key Characteristics:
1. Dynamic Environments: Anomaly detection is crucial in dynamic storage environments where operational conditions, user behaviors, and data patterns can evolve over time.
2. Data-Intensive Nature: In storage systems dealing with vast amounts of data, anomalies may signify potential inefficiencies, errors, or security breaches.
3. Real-Time Responsiveness: The ability to detect anomalies in real-time ensures timely responses to emerging issues, minimizing the risk of downtime and optimizing storage performance.
Application in Storage Monitoring:
1. Diverse Storage Metrics:
- Anomaly detection in our storage monitoring system involves continuously monitoring a range of diverse metrics, including 'eventType,' 'detailType,' , 'isDirectory' and ‘IP address’ among others.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.8 March - April – 2024
DOI:10.46647/ijetms.2024.v08i02.020 ISSN: 2581-4621

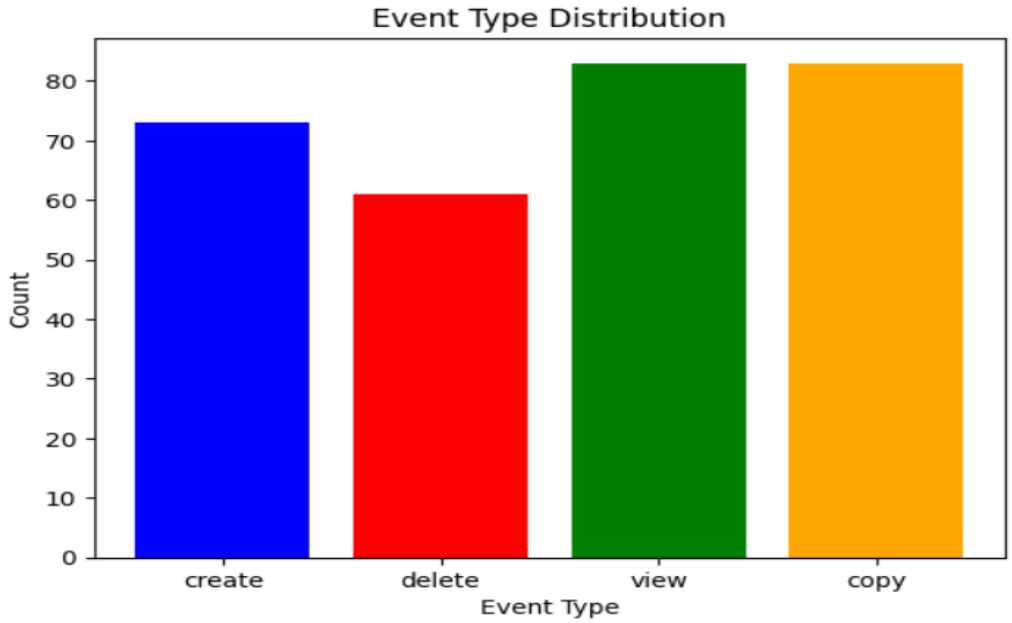
o Deviations in these metrics may indicate irregular activities, potential security threats, or operational issues.



	eventType	detailType	isDirectory
0	copy	view-file	True
1	copy	delete-file	False
2	create	create-file	True
3	view	create-file	False
4	copy	copy-file	False

2. Proactive Management:
o By employing machine learning techniques like Isolation Forest, the system can proactively identify abnormal patterns or behaviors, allowing administrators to take preemptive actions before issues escalate.
o Proactive management contributes to minimizing downtime, enhancing data integrity, and optimizing resource utilization.
3. Efficient Resource Utilization:
o Anomaly detection assists in efficiently allocating and managing storage infrastructure resources.
o By promptly identifying anomalies, administrators can optimize resource allocation, preventing underutilization or overload situations.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.8 March - April – 2024
DOI:10.46647/ijetms.2024.v08i02.020 ISSN: 2581-4621



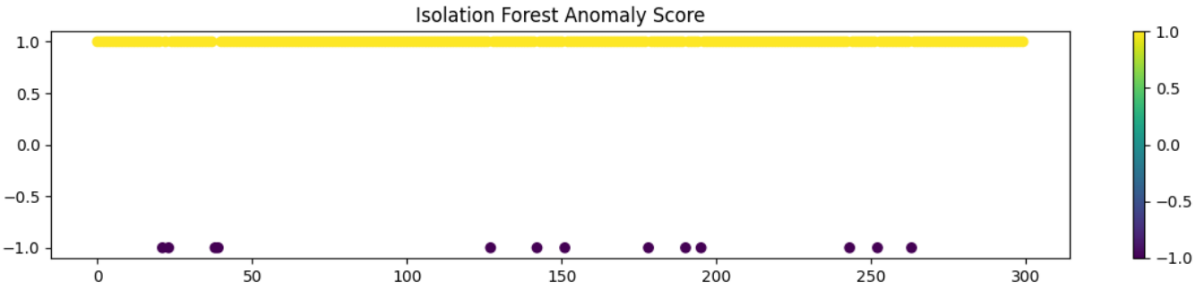
1.1.Isolation Forest:

Outliers in datasets can be efficiently identified using the isolation forest algorithm. Liu, Ting, and Zhou developed it in 2008, and it excels at handling high-dimensional data with simplicity, scalability, and efficiency. The key principle behind Isolation Forest is to isolate anomalies by constructing random decision trees and measuring the ease with which instances can be separated from the rest of the data.

How Isolation Forest Works:

1. Random Tree Construction:
   - Isolation Forest builds an ensemble of isolation trees, each constructed by recursively partitioning the data.
   - A random subset of features is chosen at each split, promoting diversity among the trees.
2. Isolation by Path Length:
   - The isolation of anomalies is measured by the average path length required to isolate them in the trees.
   - Anomalies, being less frequent, are expected to have shorter paths to isolation.
3. Scoring Anomalies:
   - Instances with shorter average path lengths across trees are considered more likely to be anomalies.
   - The anomaly score is calculated based on the normalized average path length.

As indicated by the score of -1, the Isolation Forest model identified 10 anomalies in your data. With a score of 1, the remaining 290 data points are considered normal.

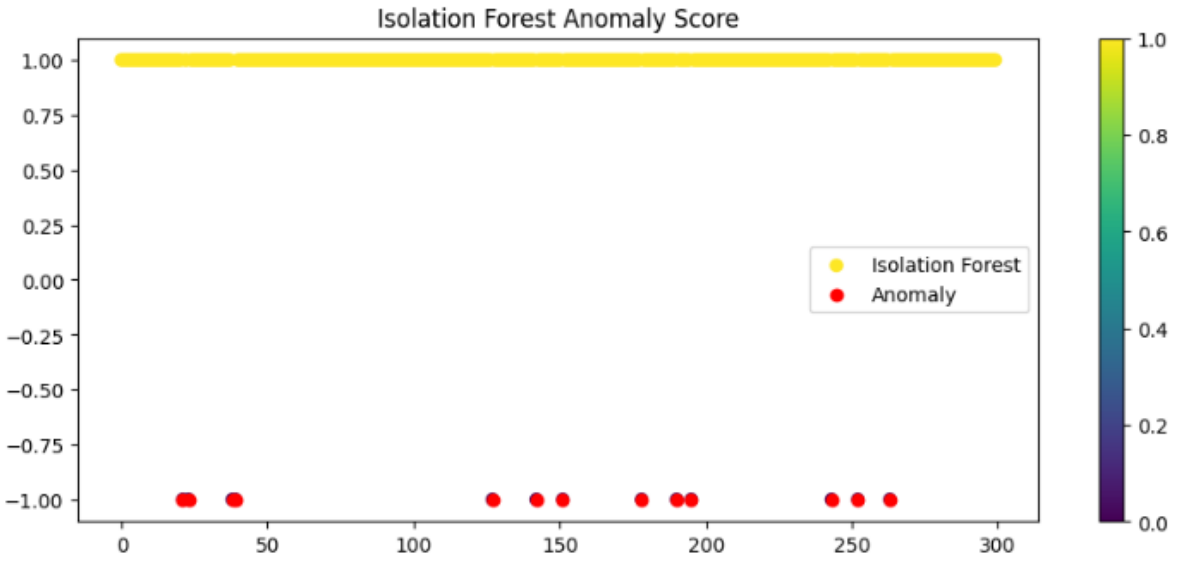


Data points representing anomalies are shown in red, while data points representing normality are shown in blue.

It uses the principle of isolating anomalies rather than profiling normal points, which is the most

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.8 March - April – 2024
DOI:10.46647/ijetms.2024.v08i02.020 ISSN: 2581-4621

common technique for unsupervised learning. If the model is correctly specified, the anomalies should correspond to the unusual patterns in your log data that might indicate a problem.



1.2.One-Class SVM

In situations where only normal instances are available during training, One-Class Support Vector Machines (One-Class SVMs) are powerful anomaly detection algorithms used to identify outliers in datasets. In high-dimensional spaces, One-Class SVM is renowned for its versatility and effectiveness. In our storage monitoring project, we leverage One-Class SVM as a key component to achieve real-time anomaly detection and proactive management within storage infrastructure.

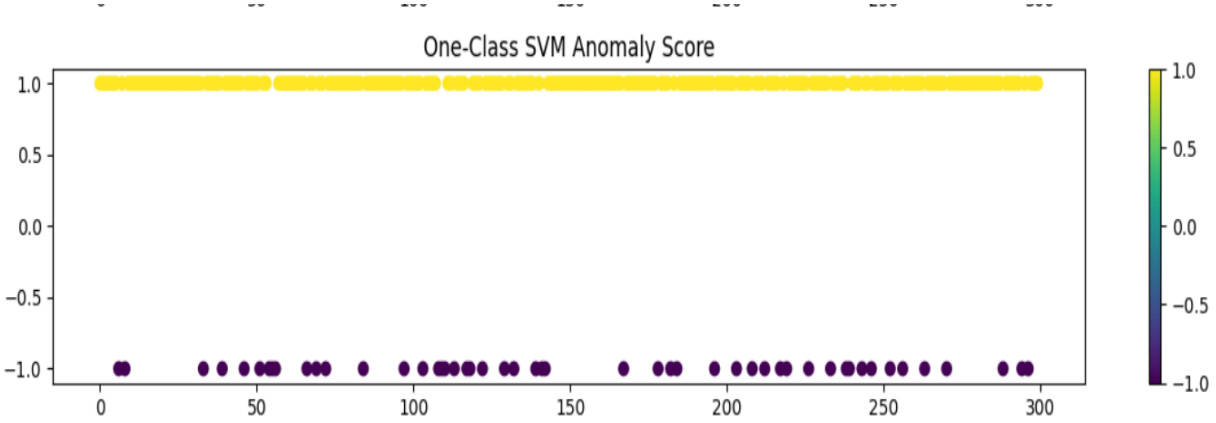
How One-Class SVM Works:

1. Single-Class Learning:
   - One-Class SVM is trained on the normal instances alone during the learning phase, making it suitable for situations where anomalies are less prevalent.
   - The algorithm learns the boundaries of the normal class, aiming to encapsulate it within a high-dimensional hyperplane.
2. Separation Hyperplane:
   - One-Class SVM constructs a separation hyperplane around the normal instances, maximizing the margin to encompass as many normal instances as possible.
   - Anomalies falling outside this hyperplane are considered outliers.
3. Kernel Trick:
   - Utilizing the kernel trick, One-Class SVM can implicitly map instances into higher-dimensional spaces, enhancing its ability to capture complex relationships within the data.

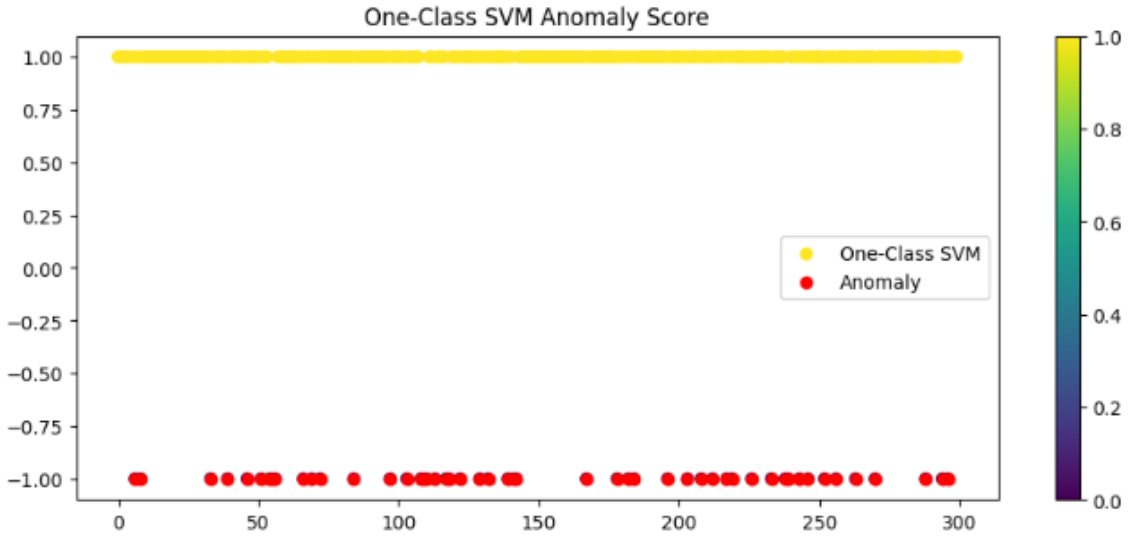
One-Class SVMs are trained on normal instances, capturing the boundaries of expected behavior. Detecting anomalies within storage metrics can be accomplished in real-time with the use of One-Class SVM, which is particularly beneficial in storage environments where anomalies are infrequent. Our system can navigate dynamic environments with One-Class SVM's adaptability to changing storage conditions, ensuring accurate anomaly detection even as operational patterns change. As part of our storage monitoring system, One-Class SVM provides a robust and versatile tool for proactive management of anomalies. In addition to minimizing downtime and optimizing storage performance, this algorithm learns the boundaries of normal behavior and identifies outliers in real-time.

153 instances have been classified as normal by the One-Class SVM model, while 147 instances have been classified as anomalous by the model (score-1).

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.8 March - April – 2024
DOI:10.46647/ijetms.2024.v08i02.020 ISSN: 2581-4621



The yellow dots at the top represent the normal data points, while the red dots at the bottom represent the anomalies.
One-Class SVM is a type of Support Vector Machine that's been adapted for use with single-class classification, or anomaly detection. It works by trying to separate the data in high-dimensional space using a hyperplane. Data points that fall on one side of the hyperplane are classified as normal, while those on the other side are classified as anomalies.



1.3.Local Outlier Factor (LOF)
A popular anomaly detection algorithm, Local Outlier Factor (LOF), assesses the relative density deviations within a dataset in order to identify outliers. It was developed by Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander in 2000. In our storage monitoring project, LOF is harnessed to provide actionable insights into potential anomalies, thereby contributing to the system's overarching objectives of proactive management and efficient resource utilization.

How Local Outlier Factor (LOF) Works:
1. Local Density Comparison:
- LOF evaluates the local density of a data point by comparing its density to that of its neighbors.
- Anomalies often exhibit lower local density compared to their neighbors, resulting in higher LOF scores.

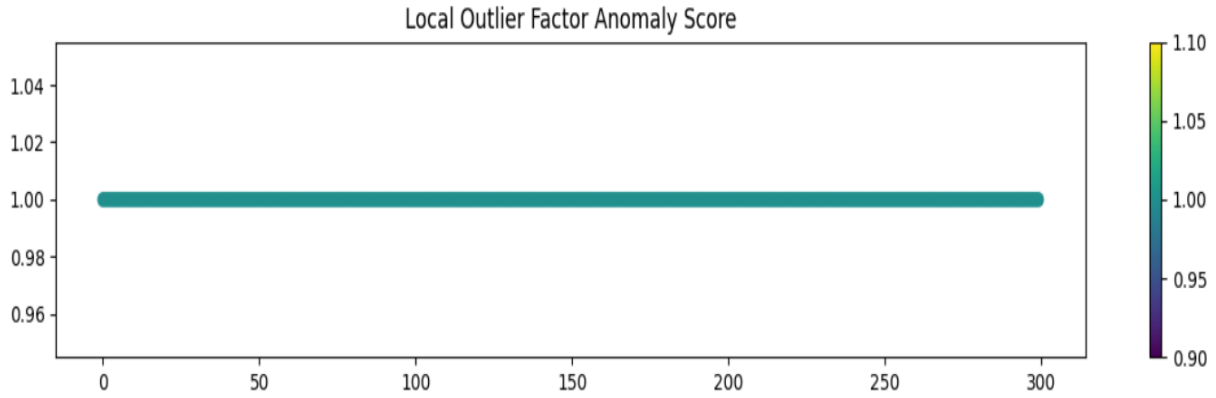
2. Neighborhood Reachability:

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.8 March - April – 2024
DOI:10.46647/ijetms.2024.v08i02.020 ISSN: 2581-4621

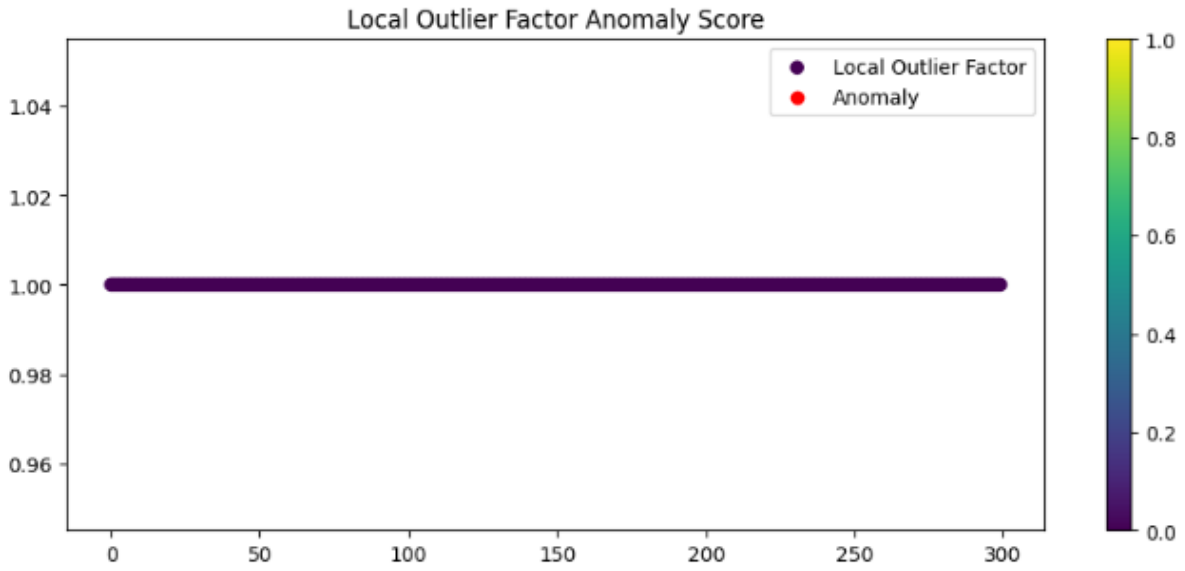
- LOF considers the reachability distance between a point and its neighbors, providing a measure of how easily a point can be reached from its neighbors.
- Outliers with lower reachability distances are assigned higher LOF scores.

3. Scoring Anomalies:

- LOF assigns a score to each data point based on the degree to which it deviates from the local density patterns.
- Higher LOF scores indicate a higher likelihood of an instance being an outlier.



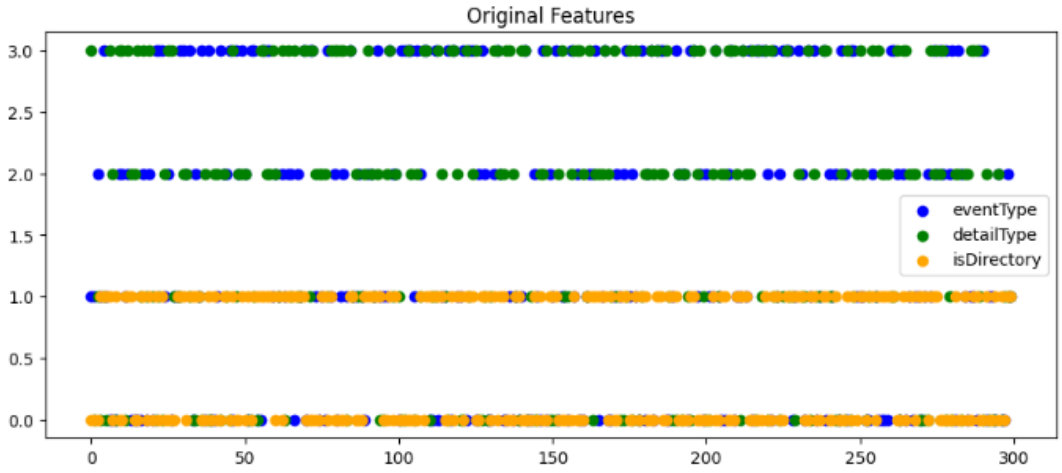
By analyzing the local density patterns of storage metrics, LFO can pinpoint instances that deviate from their local structure in order to identify anomalies. Our system detects anomalies as operational conditions change in dynamic storage environments because LOF is adaptable to varying densities. We optimize resource utilization using LOF by identifying anomalies promptly and taking proactive measures to resolve them. Local Outlier Factor (LOF) stands as a valuable tool in our storage monitoring system, contributing to the identification of anomalies exhibiting local irregularities. Its sensitivity to local structures, adaptability to varying densities, and applicability to high-dimensional data make LOF an effective algorithm for real-time anomaly detection.



International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.8 March - April – 2024
DOI:10.46647/ijetms.2024.v08i02.020 ISSN: 2581-4621

	isolation_forest_score	one_class_svm_score	local_outlier_factor_score
0	1	1	1
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
...	...	...	...
295	1	1	1
296	1	1	1
297	1	1	1
298	1	1	1
299	1	1	1

300 rows × 3 columns



Conclusion:

The Developed tool continuously monitor diverse storage metrics, identified abnormal patterns and provides actionable insights for administrators to maintain optimal storage performance, minimize downtime and enhance data integrity across various storage systems.

References:

1. V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: A survey", *ACM Comput. Surv.*, vol. 41, no. 3, pp. 71-97, 2009.
2. M. Injadat, F. Salo, A. B. Nassif, A. Essex and A. Shami, "Bayesian optimization with machine learning algorithms towards anomaly detection", *Proc. IEEE Global Commun. Conf. (GLOBECOM)*,

DOI:10.46647/ijetms.2024.v08i02.020 ISSN: 2581-4621

pp. 1-6, Dec. 2018.
3. T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth and G. Langs, Unsupervised Anomaly Detection With Generative Adversarial Networks to Guide Marker Discovery, Cham, Switzerland:Springer, vol. 10265, no. 2, 2017.
4. F. Salo, M. Injadat, A. B. Nassif, A. Shami and A. Essex, "Data mining techniques in intrusion detection systems: A systematic literature review", *IEEE Access*, vol. 6, pp. 56046-56058, 2018.
5. F. Salo, M. N. Injadat, A. Moubayed, A. B. Nassif and A. Essex, "Clustering enabled classification using ensemble feature selection for intrusion detection", *Proc. Int. Conf. Comput. Netw. Commun. (ICNC)*, pp. 276-281, 2019.