Anomaly Detection in Univariate Time Series: Rule-Based, Unsupervised, Regression and Conformal Prediction Approaches

Ajeetkumar Rai
Department of Applied Mathematics and Statistics
Stony Brook University
ajeetkumar.rai@stonybrook.edu

Abstract

Anomaly detection in time series data is very important in the finance industry. This study compares rule-based methods including quantile, generalized extreme studentized deviation, and quantile regression with unsupervised techniques such as isolation forest and one-class SVM. To improve forecast reliability, conformal forecasting is applied, which assigns statistical confidence to the detected anomalies.

1 Introduction

Investors closely monitor market movements to predict whether the market will rise or fall. Sudden behavioral changes in markets, including stocks, ETFs, and cryptocurrencies etc, can lead to significant losses. Events such as the 2008 recession or the COVID-19 market crash are examples of how sudden and rare changes can reduce wealth. Although these events are rare, their random and rare occurrence has huge implications, often referred to as tail events or outliers in statistics. Despite decades of research to predict such uncertainties, the challenge remains: can these rare events be accurately forecast before they occur? In this study, we explore various methods for anomaly detection to suit different data types and purposes. We'll investigate rule-based approaches, including quantile tests, generalized extreme studentized deviate tests, and unsupervised methods such as isolation forest and one-class SVM, and integrate them with conformal prediction for increased reliability. Additionally, we apply quantile regression to assess its performance. The effectiveness of each method is evaluated using parameters such as F1 score, precision, recall, and accuracy.

2 Literature Review

Anomaly detection is an important domain in data science and machine learning, with applications in various fields including finance, healthcare, and cybersecurity. This review synthesizes key methodologies and concepts from the existing literature, focusing on rule-based, unsupervised, and conformal prediction approaches to anomaly detection. Chandola et al. (2009) comprehensively survey anomaly detection techniques, emphasizing challenges and methodologies applicable to different data types and domains. Their work highlights the need for domain-specific adaptation of these methods to address unique challenges in different contexts effectively. Lewis (1978) laid the statistical foundation for outlier detection, providing a robust framework for identifying deviations in data. His seminal work underscores the importance of statistical rigor in anomaly detection. Quantile methods, especially quantile regression, introduced by Koenker (2005), provide an effective approach to understanding the conditional distribution of a response variable, making it an essential tool for univariate anomaly detection. Unsupervised methods, such as Isolation Forest introduced by Liu et al. (2008), leverage a tree-based structure to isolate anomalies by recursively partitioning the data. This method is particularly effective in detecting anomalies in high-dimensional data. Similarly, Tax (2001) proposed One-Class SVM, a kernel-based method designed for scenarios where most of the data belong to a single class, making it ideal for detecting rare anomalies in univariate datasets. Conformal prediction, a relative advancement, provides a framework for uncertainty quantification and reliable anomaly detection. Angelopoulos and Bates (2021) provide a gentle introduction to the principles of conformal prediction, demonstrating its versatility across a variety of applications. Mendill et al. (2022, 2023) extend the utility of conformal prediction to practical scenarios, including gas demand forecasting, and propose tools such as PUNCC, a Python library for predictive uncertainty calibration, which substantially extends the utility of conformal prediction in real-world contexts. Evaluation metrics are crucial to understanding anomaly detection methods. Powers (2011) provides a detailed description of accuracy, recall, F1 score, and ROC analysis, providing insights into selecting appropriate metrics for evaluating models across different tasks. Additionally, practical tools such as the ADTK library provide an important framework for anomaly detection, emphasizing ease of use and flexibility in integrating different methodologies. The reviewed literature outlines the development of law-based and rule-based models for anomaly detection in sophisticated machine learning and conformal prediction frameworks. Each method offers unique advantages, and their comparative performance evaluation is essential to effectively address domain-specific challenges in anomaly detection.

3 Anomaly

Anomaly is something that is not normal. Any data point that is placed at a distance from all normal data points is an anomaly. Hence anomalies are also called outliers. Anomaly detection is also called deviation detection because the attribute values of abnormal objects are different from all normal data objects. Anomaly detection methods can be classified into two main categories: point anomalies and contextual anomalies.

- *Point anomalies*: A data point is considered an anomaly because it differs substantially from the rest of the dataset, and no context is needed for the anomaly to become apparent.
- *Contextual anomalies*: A data point may appear normal on a global scale but becomes abnormal when viewed in its specific context.

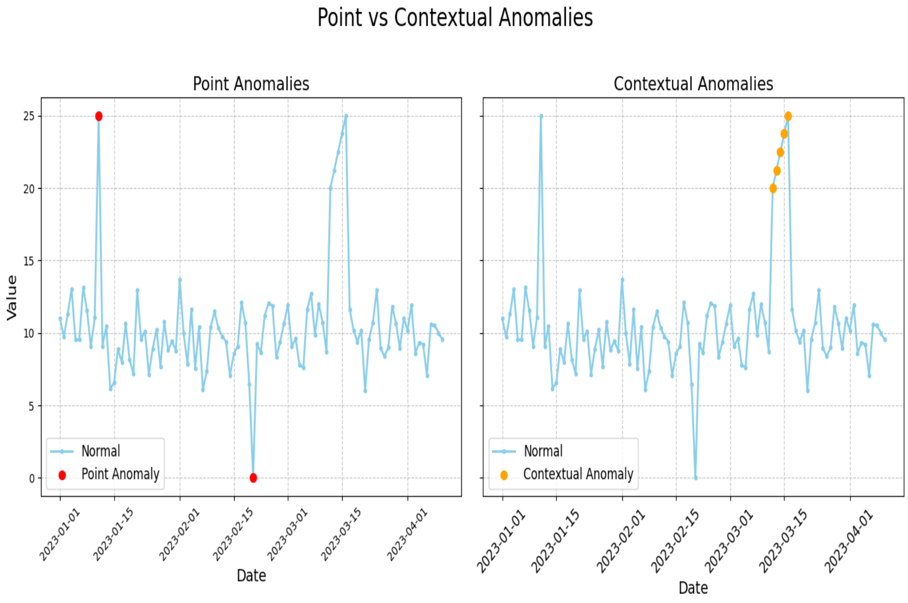

Figure 1: Point and contextual anomalies

4 Problem Statement and Dataset

When the volume of an index, which is the total number of stocks or contracts traded during a given period, is either very high or very low, it can provide important insights into market sentiment and potential price movements. High volume indicates increased trading activity and often reflects increased interest in the index or underlying securities. This can be due to major news events, announcements, or important economic data releases. High volume often confirms the strength of the price movement. Low volume indicates low trading activity and may reflect a lack of confidence among market participants. Price movements on low volume are often considered less reliable or weak trends and low-volume rallies may indicate that fewer investors are buying into the price increase, indicating that the trend may not last long. Also, if volume rises to unusually high levels, it may reflect panic selling or enthusiastic buying. Both scenarios are volatile and often precede reversals and if volume drops significantly, it could indicate that the index is losing relevance, or traders are shifting their focus elsewhere. In such scenarios, investors either lose their wealth or they will miss the chance to make a significant profit. However, these events do not occur daily and can be considered as anomalies. Hence objective is used to predict these anomalies in time series index data. The index data ranging from the year January 2020 to October 2024 has been taken for analysis. For training purposes data from January 2020 to December 2023 has been utilized and the remaining data is used for testing purposes. The volume is given in millions and if the volume is more than 500 million or less than 250 million then it is marked as an anomaly.

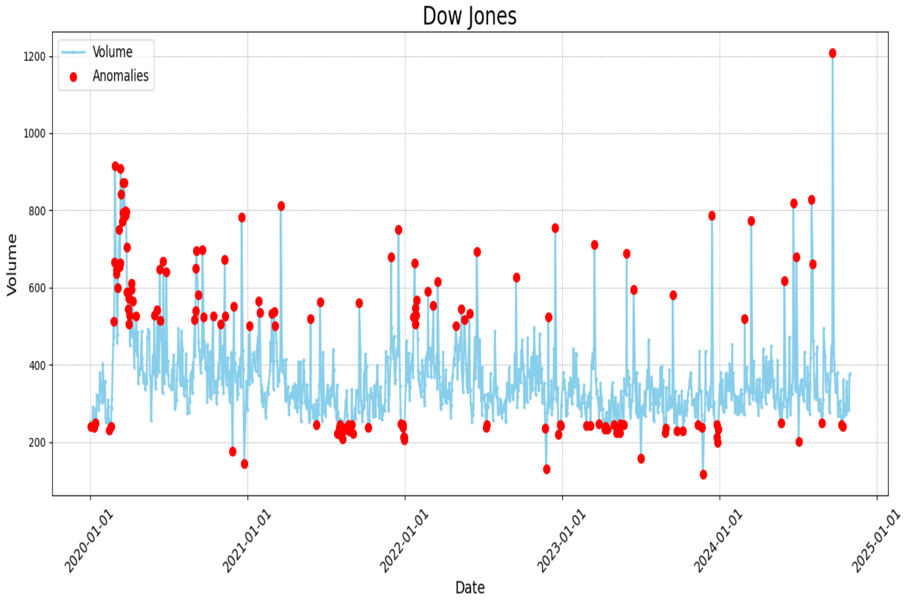

Figure 2: Dow Jones index data

5 Methodology

Various anomaly detection methods such as Quantile, Generalized Extreme Studentized Deviate, Quantile Regression, Isolation Forest, and One Class-SVM have been implemented. Conformal prediction has been implemented using the Python library PUNCC where the base model Isolation Forest and One Class-SVM are used. The model is trained on the training dataset and validated on the test data. F1 score, precision, actually, and accuracy have been used for evaluation.

5.1 Rule-based anomaly detection

In rule-based anomaly detection methods, we calculate lower and upper thresholds using statistical methods and mark the data points as anomalies if they fall outside the thresholds. Below are two rule-based methods using quantiles and the generalized extreme Studentized deviation explained.

5.1.1 Quantiles

Quantiles are cut points that divide the range of a probability distribution into continuous intervals with equal probabilities.

In this experiment, an upper threshold quantile ($\alpha = 0.9$) and a lower threshold quantile ($\alpha = 0.1$) were used. Data points falling outside these thresholds are marked as anomalies.

Quantiles are defined mathematically as the inverse of the cumulative distribution function (CDF):

$$q(\alpha) = F^{-1}(\alpha)$$

This means that a quantile $q(\alpha)$ is a solution to the equation:

$$F(x) = \alpha \quad i.e., \quad F(q(\alpha)) = \alpha$$

Equivalently, the quantile can be represented in terms of the complementary probability:

$$p(x) = 1 - \alpha \quad i.e., \quad p(q(\alpha)) = 1 - \alpha$$

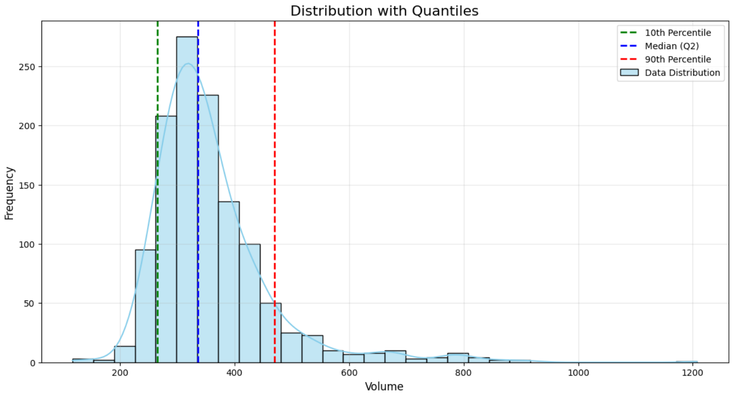

Figure 3: Volume Distribution with Quantiles

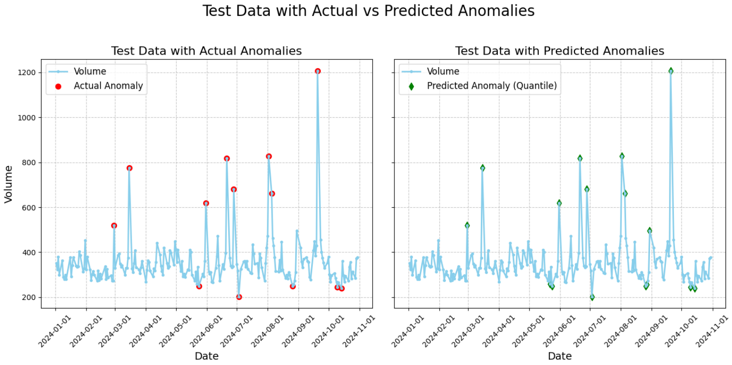

Figure 4: Quantile Thresholds: Actual and Predicted Anomalies

5.1.2 Generalized Extreme Studentized Deviate

The Generalized Extreme Studentized Deviate (GESD) Test is a statistical method used to detect outliers in a univariate dataset. This iterative test identifies one outlier at a time and removes it, recalculating the test statistic and significance level for the remaining data in each step. The data should approximately follow a normal distribution and parameters are univariate data and a number of outliers. If the test statistic R_i exceeds the critical value λ_i, the observation corresponding to R_i is considered an outlier. The procedure is iterative. After identifying an outlier, it is removed from the dataset, and the test is repeated on the remaining data to detect additional outliers. This process continues until no outliers are detected or the maximum number of suspected outliers has been evaluated.

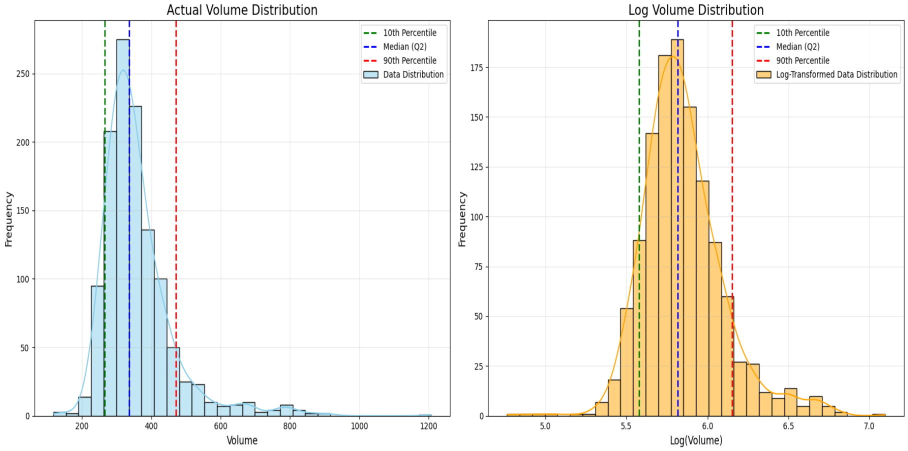

Figure 5: Original Volume and Log Volume

The test statistic R_i is computed as:

$$R_i = \frac{\max\left(|x_i - \bar{x}|\right)}{s}$$

where:

- x_i is the data point under consideration,
- $\bar{x}$ is the mean of the dataset,
- s is the standard deviation of the dataset.

The critical value λ_i is calculated as:

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i-1+t_{p,n-i-1}^2)(n-i+1)}}$$

where:

- n is the total number of observations,
- i is the index of the current suspected outlier (starting from 1),
- $t_{p,n-i-1}$ is the critical value from the t-distribution with $n-i-1$ degrees of freedom,
- $p = 1 - \frac{\alpha}{2(n-i+1)}$ is the adjusted significance level.

The procedure stops when $R_i \leq \lambda_i$, indicating that no further outliers exist in the data.

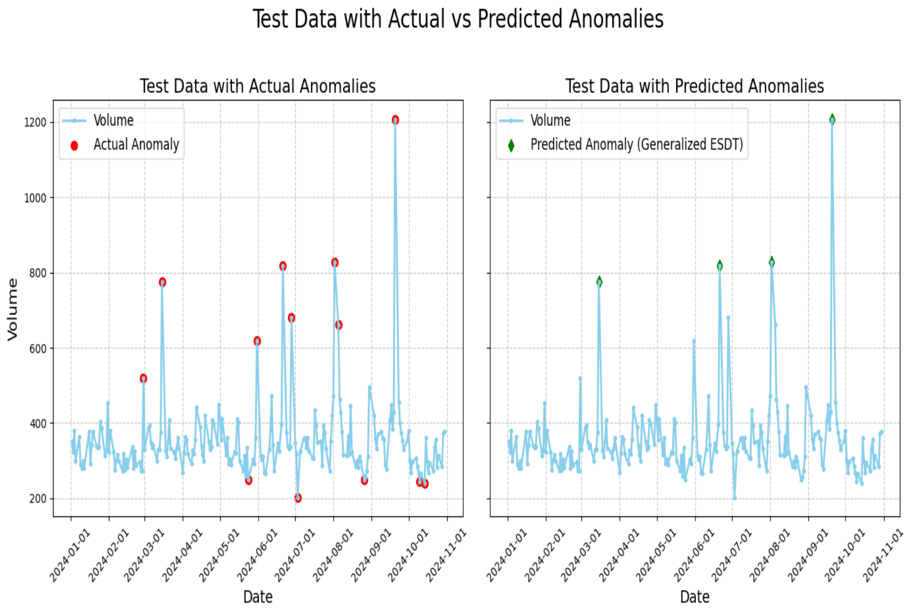

Figure 6: Generalized ESDT: Actual and predicted anomalies

Since the Generalized ESD deals with the upper tail, the lower anomalies are not detected.

5.2 Unsupervised Models

5.2.1 Isolation Forest

Isolation Forest is an algorithm for data anomaly detection using binary trees. It is based on the assumption that because anomalies are few and different from other data, they can be isolated using few partitions.

Anomaly detection using the Isolation Forest algorithm is performed as follows:

1. Use the training dataset to build a number of Isolation Trees (iTrees).
2. For each data point in the test set:
   (a) Pass it through all the iTrees, counting the path length for each tree.
   (b) Assign an *anomaly score* to the instance.
   (c) Label the point as an *anomaly* if its score exceeds a predefined threshold, which depends on the domain.

The algorithm for computing the anomaly score of a data point is based on the observation that the structure of iTrees is equivalent to that of Binary Search Trees (BSTs): a termination to an external node of an iTree corresponds to an unsuccessful search in the BST. Therefore, the estimation of the average path length $h(x)$ for external node terminations is the same as that of the unsuccessful searches in BSTs, given by:

$$c(m) = \{\, 2\; H(m-1) - \frac{2(m-1)}{n}, for\, m > 2, 1, for\, m = 2, 0, otherwise.$$

where:

- n is the test set size,
- m is the sample set size,
- $H(i)$ is the harmonic number, which can be approximated as:

  $$H(i) = \ln(i) + \gamma$$

  with $\gamma = 0.5772156649$ being the Euler-Mascheroni constant.

Above, $c(m)$ is the average $h(x)$ given m. This can be used to normalize $h(x)$ and compute the anomaly score for a given instance x as:

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

where $E(h(x))$ is the average path length $h(x)$ from a collection of iTrees. For any data point x:

- If $s(x, m) \approx 1$, then x is very likely an anomaly.
- If $s(x, m) < 0.5$, then x is likely normal.
- If all points in the sample score are around 0.5, then it is likely that all points are normal.

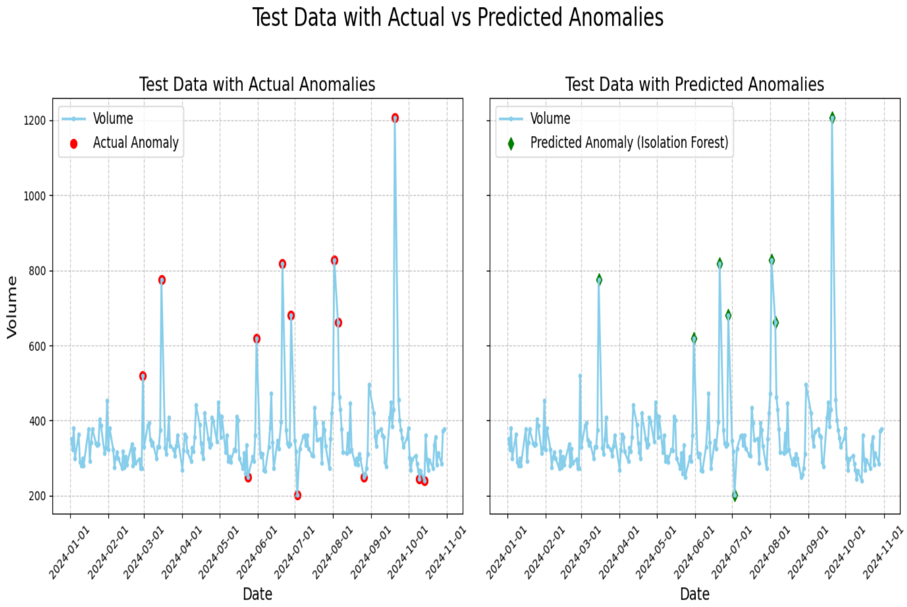

Figure 7: Isolation Forest: Actual and predicted anomalies

5.2.2 One Class - Support Vector Machine

SVM-based one-class classification (OCC) relies on identifying the smallest hypersphere, with radius r and center c, consisting of all the data points. This method is known as Support Vector Data Description (SVDD).

The original problem can be defined as:

$$\min_{r,c} r^2 \quad subjectto \quad \|\Phi(x_i) - c\|^2 \leq r^2 \quad \forall i = 1, 2, \ldots, n$$

However, this formulation is restrictive and sensitive to outliers. To allow for the presence of outliers, a more flexible formulation is given as:

$$\min_{r,c,\zeta} r^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \zeta_i$$

$$subjectto \quad \|\Phi(x_i) - c\|^2 \leq r^2 + \zeta_i \quad \forall i = 1, 2, \ldots, n$$

Karush-Kuhn-Tucker Conditions for Optimality

From the Karush-Kuhn-Tucker conditions for optimality, we obtain:

$$c = \sum_{i=1}^{n} \alpha_i \Phi(x_i)$$

where the α_i's are the solution to the following optimization problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i \kappa(x_i, x_i) - \sum_{i,j=1}^{n} \alpha_i \alpha_j \kappa(x_i, x_j)$$

$$\sum_{i=1}^{n} \alpha_i = 1 \quad and \quad 0 \leq \alpha_i \leq \frac{1}{\nu n} \quad forall \quad i = 1, 2, \ldots, n$$

The introduction of a kernel function provides additional flexibility to the One-Class SVM (OSVM) algorithm.

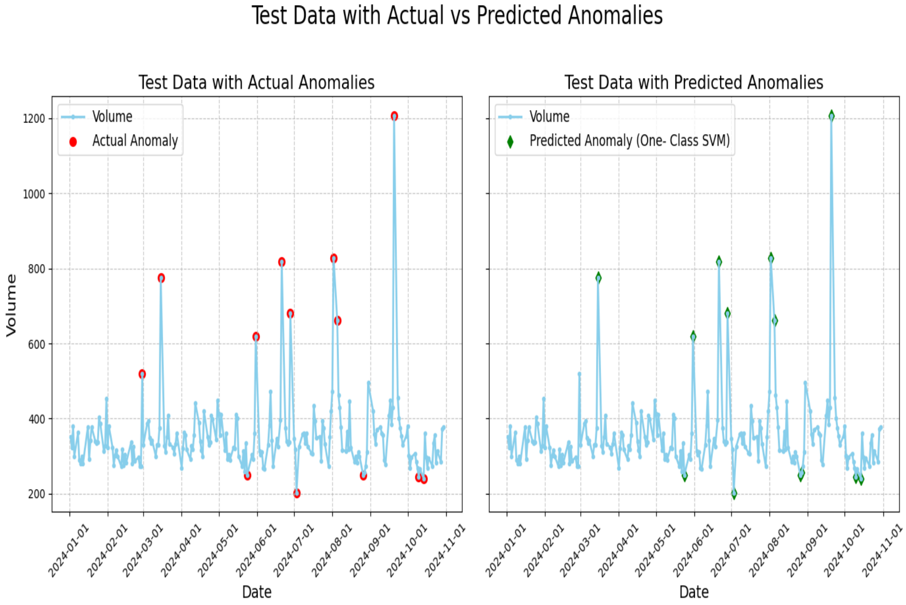

Figure 8: One Class - SVM : Actual and predicted anomalies

5.3 Regression

5.3.1 Quantile Regression

Standard regression works with the mean of the distribution and works effectively with homoscedasticity data. But, when data is heteroscedasticity standard regression fails and the alternate model is quantile regression. The idea is to implement quantile regression where upper and lower quantile is placed such that data points above and below these quantile lines will be flagged as anomalies.

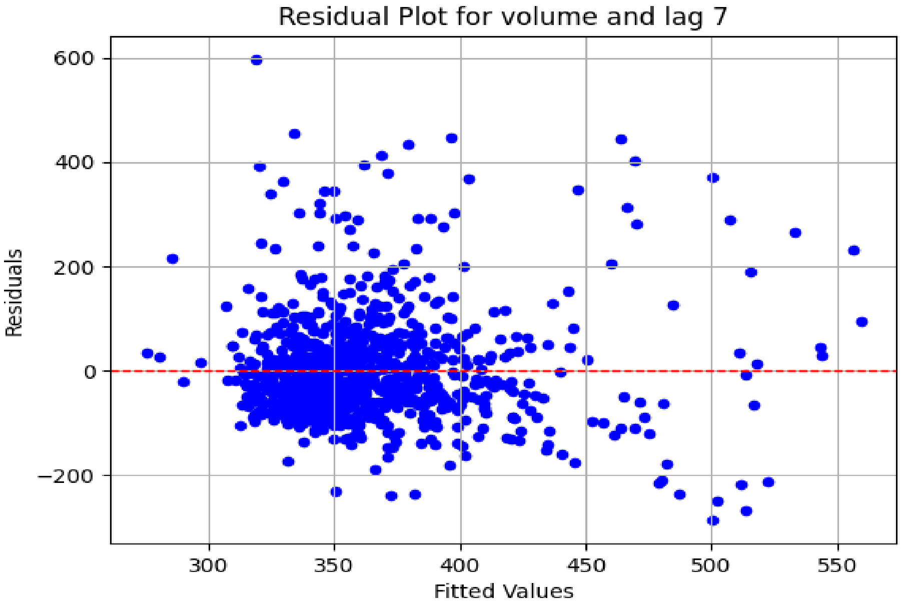

Figure 9: Weekly lags and volume

As we can see lag 7 and volume perfectly fits with the data requirement for quantile regression. To calculate the thresholds F_1 and F_2, we first need to compute the 25th and 75th percentiles (quantiles) of the data, along with the *Interquartile Range* (IQR), which is the difference between the 75th and 25th quantiles. The steps are as follows:

1. Calculate the 25th and 75th Quantiles:

$$Q_{25} = 25thpercentile, \quad Q_{75} = 75thpercentile$$

2. Compute the Interquartile Range (IQR):

$$IQR = Q_{75} - Q_{25}$$

3. Set the threshold parameter k: We define k as $\frac{1}{2}$, a constant that controls the threshold range.

4. Calculate the Lower Threshold F_1:

$$F_1 = Q_{25} - k \times IQR$$

where $k = \frac{1}{2}$.

5. Calculate the Upper Threshold F_2:

$$F_2 = Q_{75} + k \times IQR$$

These thresholds are used for anomaly detection, where data points below F_1 or above F_2 are considered anomalies.

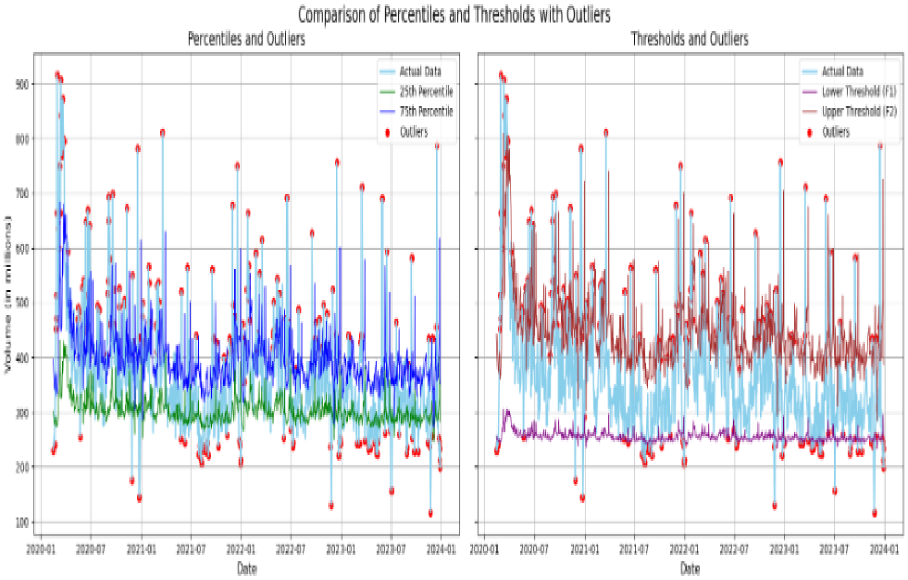

Figure 10: Quantile Regression (Train): Actual and predicted anomalies

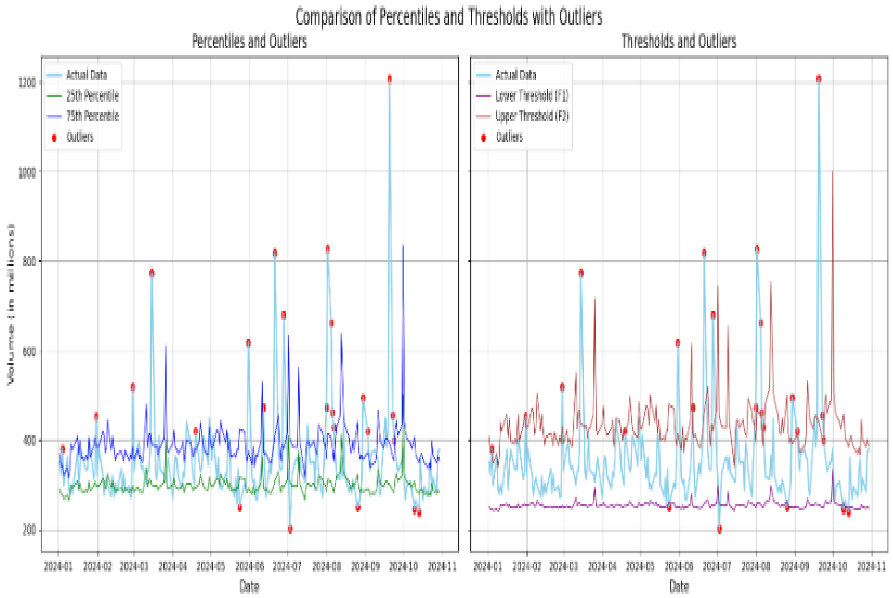

Figure 11: Quantile Regression (Test): Actual and predicted anomalies

5.4 Conformal Anomaly Detection

Conformal prediction is a framework that provides a way to create prediction sets based on past data. These sets are associated with a confidence level, meaning that with a certain probability, the true outcome will lie within the predicted set.

Conformal prediction can be extended to handle unsupervised anomaly detection, allowing us to identify data points that do not conform to the ”normal” (or nominal) distribution of a dataset. The goal is to assign a statistical guarantee to the anomaly detector, ensuring control over the false positive rate.

To detect anomalies, we start with a model that assigns an anomaly score $S(x)$ to each data point X_i. Higher scores indicate a higher likelihood of being an outlier.

For each example x_i in the calibration dataset, compute the nonconformity score as $R_i = S(x_i)$, and store all nonconformity scores in a vector R.

Next, compute the anomaly score threshold δ_α as the $(1-\alpha)(1+1/n_{calib})$-th empirical quantile of R.

For a new test point X_{new}, the conformalized anomaly detector classifies it as:

$$\bar{C}_\alpha = \{ \; Normal, if s(X_{new}) \leq \delta_\alpha, Anomaly, otherwise.$$

Conformal anomaly detection provides an error control guarantee, meaning that under the assumption of exchangeability, the probability of a false positive (labeling a nominal instance as an anomaly) is bounded by α.

Conformal anomaly detection is implemented with the unsupervised models Isolation Forest and one class - SVM, as described above.

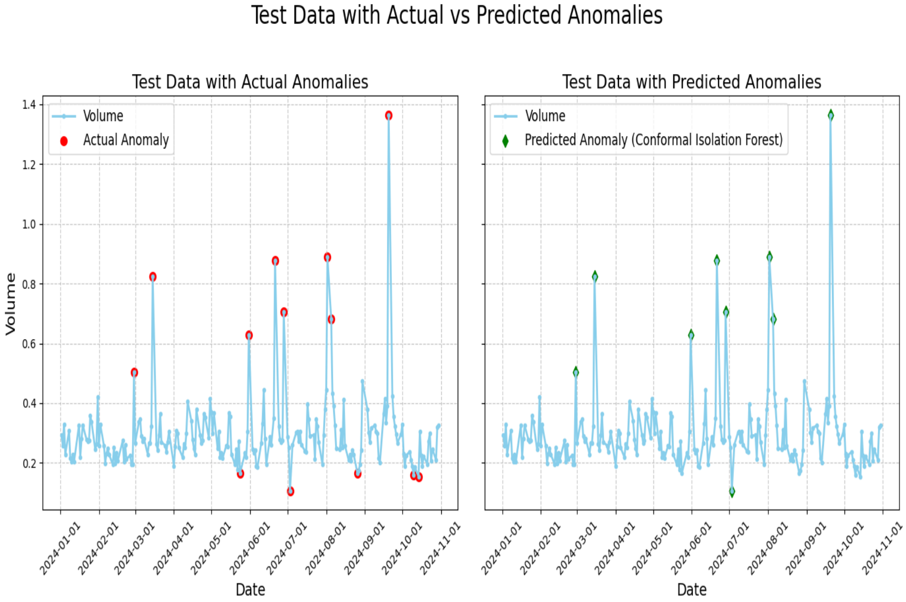

Figure 12: Conformal Isolation Forest : Actual and predicted anomalies

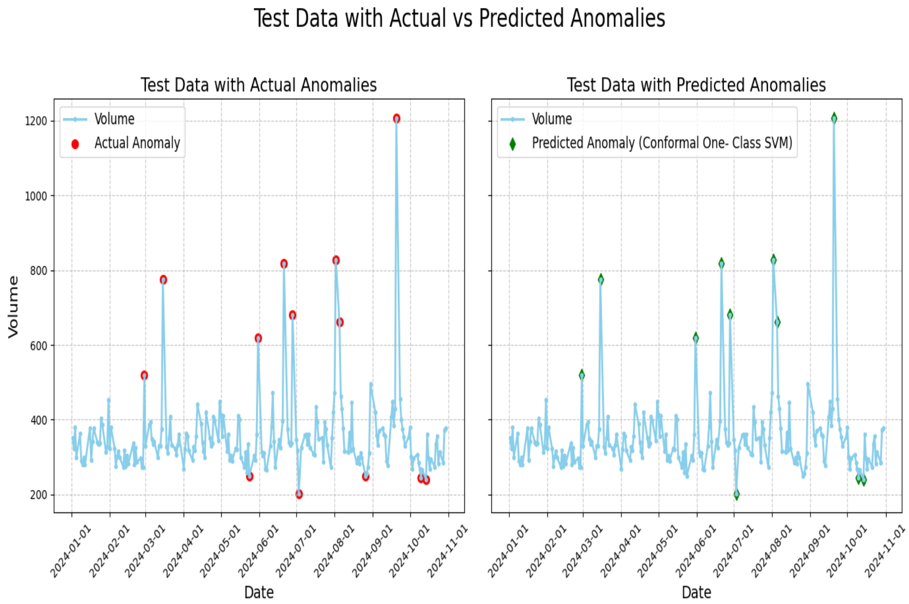

Figure 13: Conformal One Class - SVM : Actual and predicted anomalies

6 Metrics and Performance

In the table below we can see how our model is performing.

Table 1: Anomaly Detection Method Evaluation

Model	Type	Data	F1 Score	Precision	Recall	Accuracy
ESD	Rule-based	Test	0.47	1.00	0.31	0.96
ESD	Rule-based	Train	0.21	1.00	0.12	0.87
Quantile	Rule-based	Test	0.90	0.81	1.00	0.99
Quantile	Rule-based	Train	0.84	0.72	1.00	0.94
Isolation Forest	Unsupervised	Test	0.76	1.00	0.62	0.98
Isolation Forest	Unsupervised	Train	0.77	1.00	0.63	0.95
Isolation Forest (Conformal)	Conformal	Test	0.82	1.00	0.69	0.98
Isolation Forest (Conformal)	Conformal	Train	0.67	1.00	0.50	0.93
One Class SVM	Unsupervised	Test	0.92	0.92	0.92	0.99
One Class SVM	Unsupervised	Train	0.83	0.81	0.85	0.95
One Class SVM (Conformal)	Conformal	Test	0.92	1.00	0.85	0.99
One Class SVM (Conformal)	Conformal	Train	0.99	1.00	0.97	1.00
Quantile Regression	Regression	Test	0.58	0.41	1.00	0.91
Quantile Regression	Regression	Train	0.72	0.60	0.91	0.90

The below plot shows how different models work in training and test data using different metrics like F1 score, precision, recall, and accuracy.

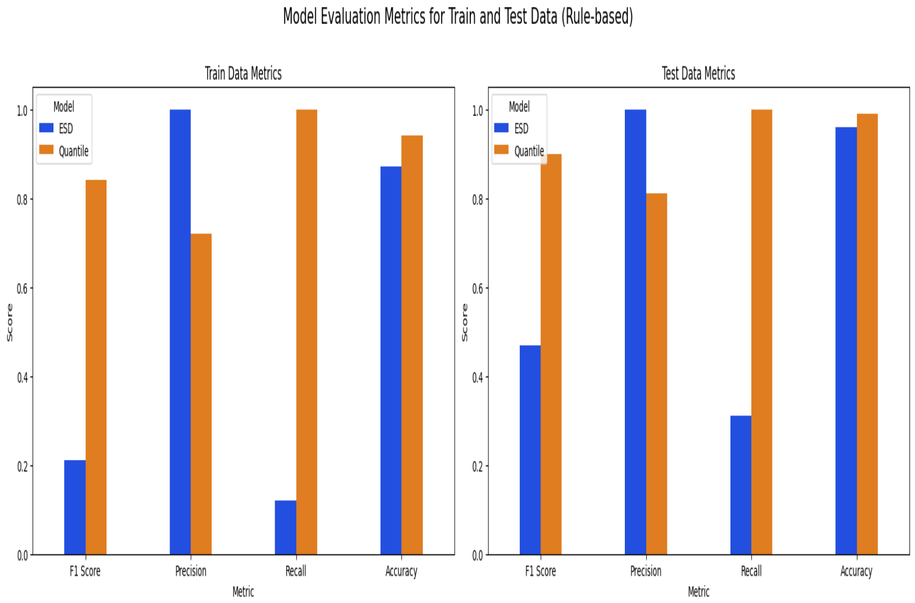

Figure 14: Rule-Based Models: Evaluation metrics

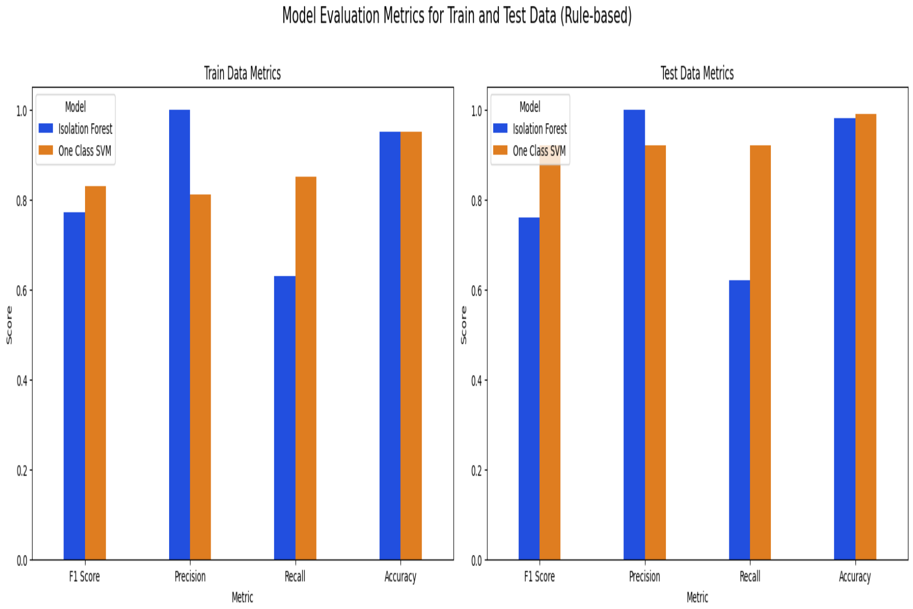

Figure 15: Unsupervised Models: Evaluation metrics

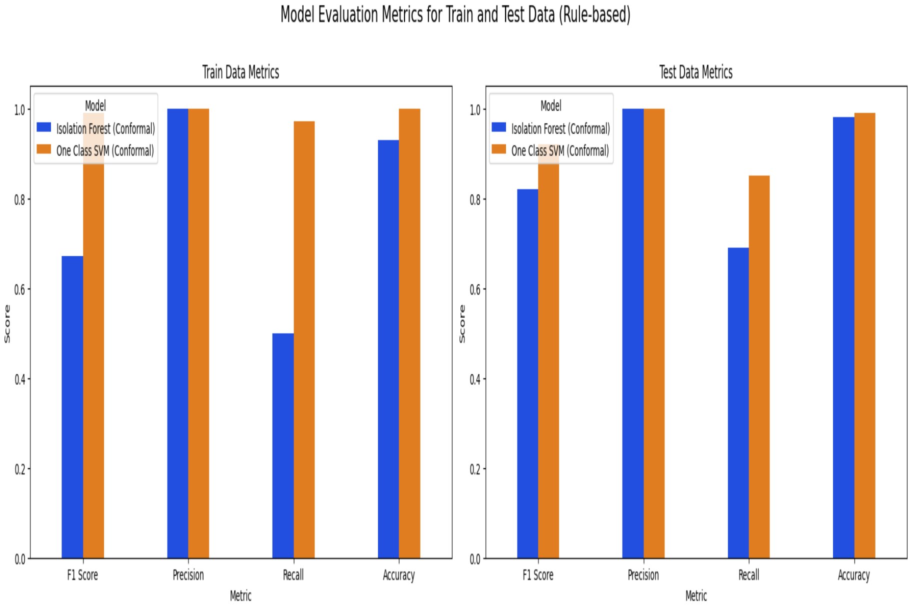

Figure 16: Conformal Models: Evaluation metrics

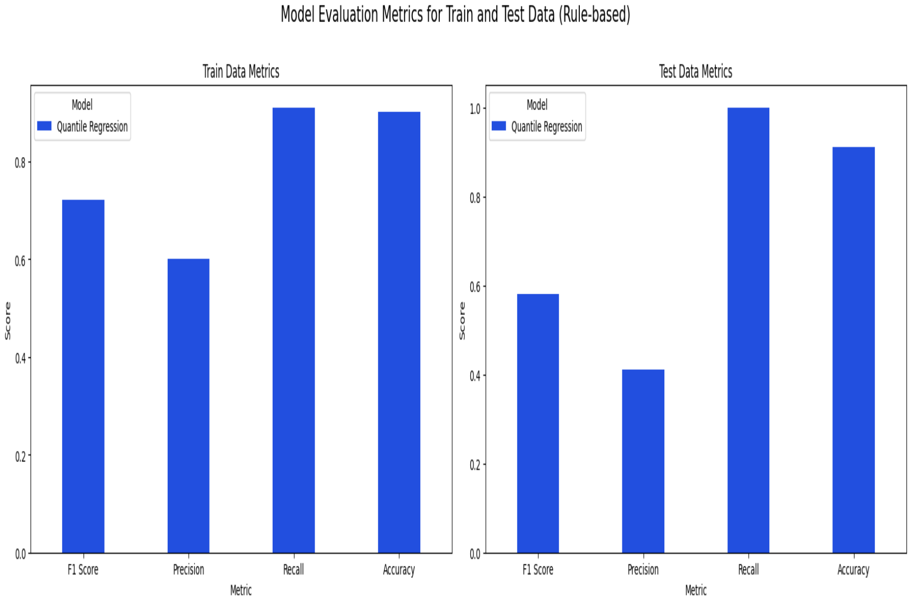

Figure 17: Regression Model: Evaluation metrics

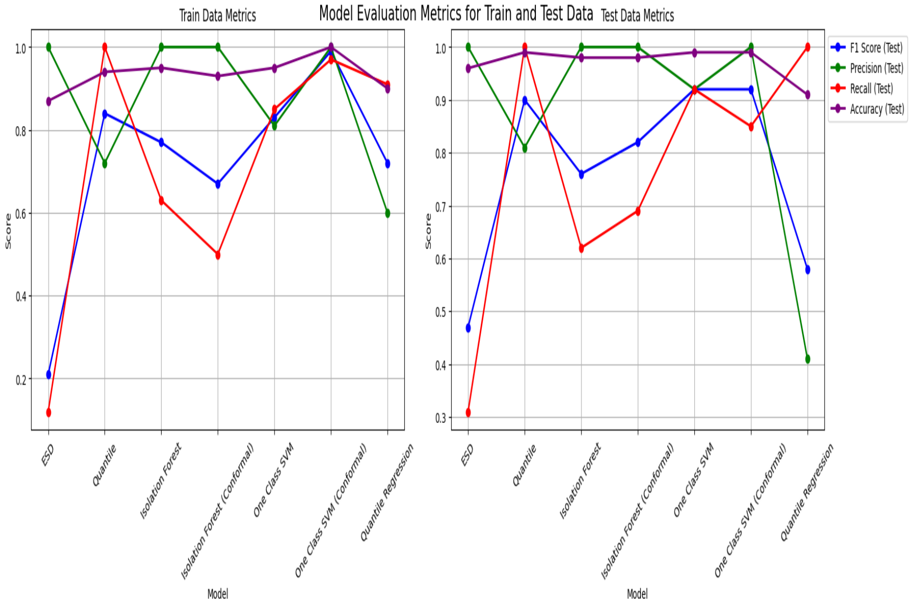

Figure 18: Evaluation metrics

7 Conclusion

The generalized extreme studentized deviation (ESD) test performs poorly, which is expected since it mainly focuses on detecting anomalies in the upper tail of the distribution. In contrast, the quantile method exhibits strong performance, as it effectively identifies anomalies by considering volumes that are greater than 600 million or less than 250 million. However, Isolation Forest and its conformal version do not perform well in this context. On the other hand, One-Class SVM and its conformal version outperform all other models, demonstrating better anomaly detection capabilities. Finally, quantile regression also performs well, providing a creative and effective approach to characterizing anomalies.

References

1. Chandola, Varun Banerjee, Arindam Kumar, Vipin. (2009). Anomaly Detection: A Survey. ACM Comput. Surv.. 41. 10.1145/1541880.1541882.

2. Barnett, V., Lewis, T. (1978). Outliers in statistical data. Wiley. https://archive.org/details/outliersinstatis0000barn.

3. F. T. Liu, K. M. Ting and Z. -H. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp.

413-422, doi: 10.1109/ICDM.2008.17.

4. Tax, DMJ. (2001). One-class classification; concept-learning in the absence of counter-examples. [Dissertation (TU Delft), Delft University of Technology].

5. Angelopoulos, Anastasios; Bates, Stephen (2021). ”A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification”. arXiv:2107.07511.

6. @inproceedingsmendil2023puncc,title=PUNCC: a Python Library for Predictive Uncertainty Calibration and Conformalization,author=Mendil, Mouhcine and Mossina, Luca and Vigouroux, David,booktitle=Conformal and Probabilistic Prediction with Applications, pages=582–601,year=2023,organization=PMLR

7. @inproceedingsmendil2022robust,title=Robust Gas Demand Forecasting With Conformal Prediction,author=Mendil, Mouhcine and Mossina, Luca and Nabhan, Marc and Pasini, Kevin, booktitle=Conformal and Probabilistic Prediction with Applications,pages=169–187, year=2022,organization=PMLR

8. Koenker, Roger (2005). Quantile Regression. Cambridge University Press. pp. 146–7. ISBN 978-0-521-60827-5.

9. https://adtk.readthedocs.io/en/stable/quickstart.html

10. Powers, David M. W. (2011). ”Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness Correlation”. Journal of Machine Learning Technologies. 2 (1): 37–63. S2CID 55767944.

11. https://github.com/valeman/awesome-conformal-prediction

12. https://www.mathworks.com/help/stats/outlier-detection-using-quantile-regression.html

13. https://medium.com/aimonks/univariate-time-series-anomaly-detection-capturing-the-unusual-a5cc5ee24462.

14. https://www.analyticsvidhya.com/blog/2022/05/an-end-to-end-guide-on-anomaly-detection/

15. https://medium.com/@ngiengkianyew/quantile-regression-62a398186f7d