NU
NIIT UNIVERSITY
THE UNIVERSITY OF THE FUTURE
Industry-linked | Technology-based | Research-driven | Seamless

==========

==========

- O

CERTIFICATE BY SUPERVISOR(S)

This is to certify that the present R&D project entitled **Time Series Anomaly detection** being submitted to NIIT University, Neemrana, in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology, in the area of BT/CSE/ECE/GIS, embodies faithful record of original research carried out by **Prashant Pandey** and **Utkarsh Sharma.** They have worked under my/our guidance and supervision and that this work has not been submitted, in part or full, for any other degree or diploma of NIIT or any other University.

Place: Neemrana

Name of the Supervisor(s) with signature

Prof.Sudip Sanyal

Date: 29th May 2023

==========

DECLARATION BY STUDENT(S)

I/We hereby declare that the project report entitled **Time Series Anomaly detection** which is being submitted for the partial fulfillment of the Degree of Bachelor of Technology, at NIIT University, Neemrana, is an authentic record of my/our original work under the guidance of **Prof. Sudip Sanyal**. Due acknowledgements have been given in the project report to all other related work used. This has previously not formed the basis for the award of any degree, diploma, associate/fellowship or any other similar title or recognition in NIIT University or elsewhere.

Place: Neemrana

Date: 29th May 2023

Prashant Pandey	**BT20HCS096**	**Data Science**	**Signature**
Utkarsh Sharma	**BT20HCS214**	**CSE**	**Signature**

R & D Project Report

Academic Year- 2022-23

On

TIME SERIES ANOMALY DETECTION

Submitted by

1. Prashant Pandey	BT20HCS096	Data Science
2. Utkarsh Sharma	BT20HCS214	CSE

Prof. Sudip Sanyal

Problem Statement:

To detect all the anomalies in a given time series data

Abstract:

Time series anomaly detection can be done in many ways but here we are trying to find the best way to detect an anomaly in the time series data. We have used two algorithms to detect time series anomalies, one algorithm is a forecasting algorithm (ARIMA MODEL) and the other one is statistical algorithm (Median absolute deviation). In the ARIMA model we try to predict the future data points and compare it with the actual value to identify the point as an anomaly. In Median Absolute deviation we use the median to find how much the current value deviates from Median.

Keywords:

Auto Regression, Moving Average, Normal distribution, Standard Deviation, threshold, trends, seasonality

1. Introduction:

Time series data plays a crucial role in our lives and any anomaly in that data can be very harmful to the data processing. To detect all those anomalies we have various kinds of methods, most commonly used are forecasting and statistical models that we will be discussing in this paper further.

A forecasting method uses prediction mathematical techniques to predict the future values and then match those values distance from the predicted values. If that predicted value is very far from the actual values then that point can be said to be an anomaly.

A statistical method will more heavily rely on the statistical factors to check for an anomaly. We will be using Mean, median and mode to see the deviation of the actual values from the statistical terms.

1.1 Report Structure:

This report will start from the introduction which will give us the understanding of the problem as well as a glimpse into the solution that we have derived to solve the problem. After the introduction the related works in the same field will be shared to give us the current progress on how the problem is treated and what are some of the challenges faced by the current methods. Then there will be a detailed description of the methodology discussed which is used to solve the problem. Results will be presented after the proposed methodology which later will be followed by conclusion and future scope.

2. Related Work:

2.1 Research Paper [1]

2.1.1 Summary:

This paper by Mohammad Braei and Sebastian Wagner they talk about the best state of the art anomaly detection techniques. There are an increasing number of Machine learning algorithms being developed to detect time series anomalies. This paper discusses the basic terms required for the understanding of anomaly detection and then starts with all the models for detection of anomaly. First it discusses all the forecasting methods then with all the machine learning techniques. At last it collects all the algorithms and tells us which algorithm is used when and how.

2.1.2 Conclusion:

The authors did not take into account the hardware limitations of normal systems and resources allocation for huge machine learning algorithms. Before applying machine learning algorithms we need a huge amount of data to train the algorithms and then a very decent hardware to run those algorithms. All those machine learning algorithms are very effective but they cannot be used everywhere and an easy to use algorithm is lacking in this paper.

2.2 Research Work [2]:

This paper talks about an advanced statistical model Seasonal ESD(Extreme studentized deviate) and Hybrid Seasonal ESD (H-S-ESD). This paper then compares the Grubbs test and other statistical methods. It later discusses the implementation of statistical algorithms on the seasonal

data and how to handle seasonality. Then compares the results of different methods on all the datasets used and efficiency is calculated.

2.2.2 Conclusion:

All these algorithms discussed are very knowledge intensive and hard to understand by a normal person. Specialization in the respective field will only make these algorithms usable otherwise it would be very difficult to interpret the results in using ESD's model. Both the algorithms used are very CPU intensive which again makes it resource intensive.

2.3 Research Work[3]:

2.3.1 Summary:

This paper is written by Viacheslav Konzitsin, Iurii Katser and Dmitry Lakontsev where it use ARIMA model for forecasting of time series data and tell the step of ARIMA modeling how to do that they shown the step of autoregression , hypothesis testing and many statistical term which is using in this ARIMA modeling . In this research paper basically they have used a mathematical formula that shows how we can use an ARIMA model in online and offline ways . This is the real time series of data where the performance metric for the forecasting problem can be solved . They have told us about different algorithms which will be helpful in ARIMA modeling . In this literature they also talked about some algorithm, by which we can find the ARIMA model . in this literature they have told about some python libraries which will be useful for ARIMA modeling .

2.3.2 Conclusion:

Many scientists are developing high-performance forecasting and anomaly detection algorithms that can continuously learn and adjust to changing conditions.
In this research, we suggested a brand-new algorithm that can handle all of these difficult problems.

simultaneously. It is based on the well-known ARIMA model, which is frequently used for time-series modeling in applications of science Unsupervised problem solving is achievable thanks to the suggested algorithm's online operation and utilization of unlabeled data. In contrast to traditional ARIMA-based algorithms, it is also far more computationally efficient and almost as accurate. Online process monitoring requires each of these elements, including technical system

diagnostics. an assortment of Python3 libraries,Another suggestion made was to use the "ARIMA FD" algorithm. As a result, it can be used for both forecasting and anomaly identification.

2.4 Research Work[4]:

2.4.1 Summary:

In this literature it is basically focused on the anomaly detection of social media using the ARIMA model . They have taken the dataset of social media through different different platforms . This thesis uses time series analysis to investigate whether there has been an increase in the fixation of a radical topic1 on an online forum. This series can be seen as modeling how a person communicates on a social media platform because the variable under investigation will characterize the level of user engagement over time. When a suitable model for the series is found, it can be used to describe a potential user pattern or, more technically, the internal organization of the time series. Forecasts about the forthcoming posts' intensity can be made using the model.

However, if the actual data deviates from the pattern predicted by the model, this will be regarded as anomalous2 and can point to a change in communication, which may point to a change in the users' conduct.

Unexpected spurts of activity, which may indicate that a person on the Internet is getting increasingly hooked on a certain topic, will be considered outliers rather than a change in behavior.

Lastly, the precision of The forecasts of the model, the feasibility of a generalized data independent model, and the suitability of these methods for practical applications will be evaluated.

2.4.2 Conclusion:

When constructing short-term forecasts from several users, the ARIMA models appear to capture the data inside the confidence intervals in the majority of cases. To be more certain of accuracy for practical applications, it is preferable to only make one step ahead forecasts. Since there rarely appears to be a seasonal pattern,and over time, non-seasonal projections tend to converge on the mean value.

2.5 Research Work[5]:

2.5.1 Summary:

This literature is written by Hyeyoung Park where it was trying to make us understand about analysis of anomaly detection methods for streaming data . Outlier has previously been well

defined by many researchers. Since the type of data and anomaly has a significant impact on the effectiveness of anomaly detection systems, this study will clarify the many forms of anomalies rather than listing those definitions. Outliers include anomalous data points and noise that marginally deviates from the average but is not intriguing enough to be handled differently. On the other hand, an anomaly identifies the data points that might be relevant since they strongly deviate from other data points. One may classify anomalies as a particular kind of outlier. The current challenge in anomaly detection is to distinguish between noise and anomalies.

2.5.2 Conclusion :

There is no one perfect method that can identify all varieties of anomalies in reality, despite the fact that several unsupervised models have been developed and are in use to examine anomalies. In this research, many strategies were examined and accessed based on the sorts of abnormalities.

The prediction-based model employing ARIMA is one of the frequently used techniques to identify contextual anomalies. Actual data points and forecasted values are compared to form the foundation of ARIMA models. A data point is considered an anomaly when its actual value exceeds or falls short of an expected value and threshold. The R package “tsoutliers” was chosen in order to put the ARIMA model implementation approach into practice. Though it performs reasonably well, additional tests with In the upcoming job, the time series with numerous seasonalities must be completed. The HOT SAX method can be used to find collective abnormalities. The concept for this technique came from the discretization of time series data. Although the HOT SAX algorithm's output is very dependent on the window's length, once the user determines the right window size, it produces results that are sufficient. The R package “jmotif” contains an executable for the HOT SAX algorithm.

3. Proposed Methodology:

3.1 Workflow:

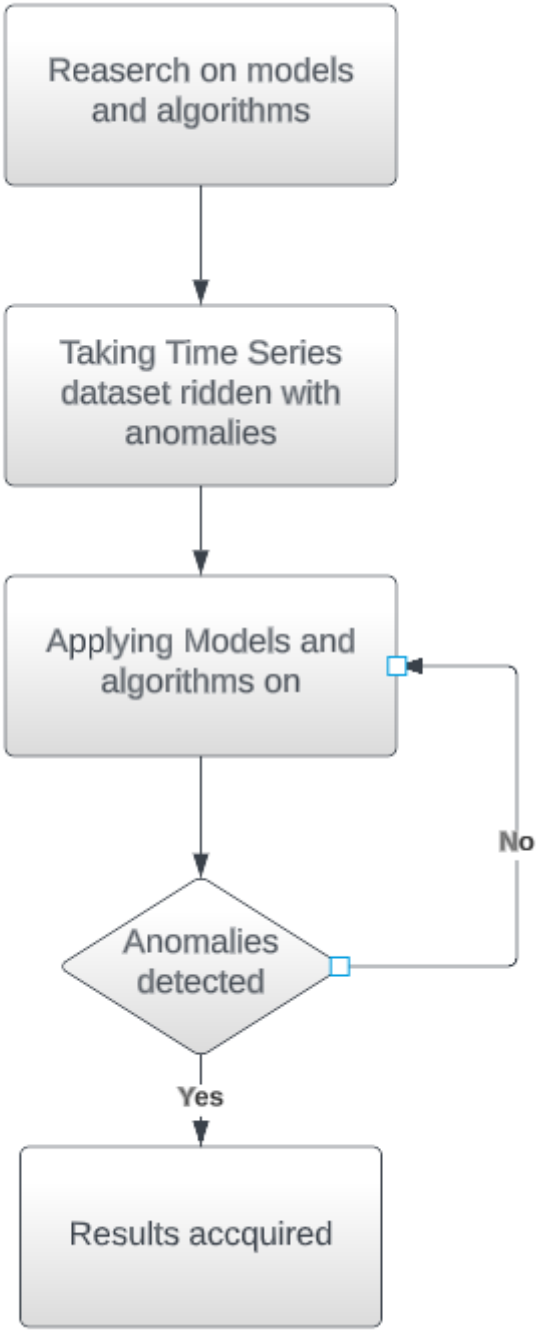

Figure1: Workflow

3.1.1 Research on Algorithms and Models:

We started with the basic research on algorithms and models which can be used and are very prominent nowadays. This provided us a basic building block to develop our understanding of the problem statement and how this problem can be solved. We came to know about many algorithms such as moving average, weighted moving average, ARIMA model etc, which are all forecasting techniques. But later it was seen that we also need some more statistical approach which uses

statistical tools like mean, median and mode, standard deviation, variance etc. Then we came to know about MAD (median absolute deviation).

3.1.2 Taking Time Series dataset ridden with anomalies:

Now after finding algorithms it was the time to find an anomaly ridden dataset to test our algorithms for their accuracy. We took the time series data in which we were already aware of the anomalies so that we can test the accuracy of the algorithms that we have selected.

3.1.3 Applying the Models and algorithms:

We applied our selected model and algorithm to the time series data that we have collected, so that we can know to what extent time series anomalies are detected and what are the problems faced by our selected algorithms. If the anomalies were not detected we applied some other algorithms and if anomalies were detected, results were saved. This process kept on repeating until we found our algorithms of 2 different natures one forecasting and other statistical profiling (ARIMA and MAD).

3.2 Technology:

3.2.1 ARIMA(Autoregression Integrated Moving Average):

3.2.1.1 Introduction:

ARIMA has various industries that use models in a variety of ways. The time series analysis technique known as ARIMA (AutoRegressive Integrated Moving Average) combines moving average (MA), autoregressive (AR), and differencing (I) components to estimate future values based on historical data. For spotting and comprehending patterns in time-dependent data, ARIMA models are useful.

This is because ARIMA models, a generic class of models used for forecasting time series data, are the cause. ARIMA (p, d, q), ARIMA models use differencing to convert a non-stationary time series into a stationary one, and they then extrapolate present values into the future. These models use "auto" correlations and moving averages over residual errors in the data to forecast future values.

Numerous fields, including finance, economics, weather forecasting, and demand forecasting, have effectively used ARIMA models. They offer a potent tool for deciphering and forecasting time-dependent data, empowering analysts and researchers to draw conclusions from past trends.

3.2.1.2 Working of this model:

There are basically three term in this which we need to understand :

Autoregression: The link between the present observation and a predetermined number of lag observations is modeled by the autoregressive component. It is predicated on the idea that linear combinations of a time series' previous values can be used to forecast its future values. The number of lagged observations taken into account in the model is indicated by the order of the autoregressive component, represented by "p." A greater "p" number denotes a longer recall of prior observations.

$$\hat{y}_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t$$

formula 1

Integrated: The time series data are subjected to differencing, which is used to weed out trends and other non-stationary patterns. Computing the difference between consecutive observations is the process of differencing. The number of differencing operations necessary to render the data steady is indicated by the differencing parameter, denoted by the letter "d." When a time series' statistical features, including mean and variance, remain constant across time, this is referred to as stationarity.

$$By_t = y_{t-1}$$

formula 2(a)

$$y'_t = y_t - y_{t-1} = (1 - B)y_t$$

formula 2(b)

$$y'_t = (1 - B)^d y_t$$

formula 2(c)

The first graph is shown to be a non-stationary time series and the second graph is stationary because its value is independent of the observational period.

Moving Average : The dependence between the error terms or residuals from earlier data is accounted for by the moving average component. It stands for the weighted total of previous mistake words. The number of lag residuals taken into account in the model depends on the order of the moving average component, represented by "q." The memory of prior errors is longer when "q" has a higher value.

$$\hat{y}_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

formula 3

An ARIMA model may identify both short-term and long-term patterns in the data by combining these three elements. Future values are predicted using the parameters determined by the model to have the best match to the observed data. The model's quality and the suitability of the parameter selection determine how accurate the forecasts will be.

3.2.1.3 Modules used:

Panda : Python's "Panda" package is a popular tool for analyzing and manipulating data. It offers data structures and operations to deal effectively with tabular and time series data as well as other structured data. The DataFrame, which resembles a table with rows and columns, much like a spreadsheet or a SQL table, is the library's primary data structure.

A user can conduct operations including filtering, sorting, joining, grouping, and aggregating data using Pandas DataFrame's many features. It also offers tools for addressing missing numbers, cleansing data, and converting data to desired formats.

Matplotlib : A popular Python library for plotting and producing visualizations is called Matplotlib. It offers a versatile and potent framework for producing numerous plot kinds, from straightforward line charts to intricate 3D visualizations.

Users can produce visual representations of their data using Matplotlib in order to obtain insights, convey discoveries, and enhance decision-making procedures. Users of the library have complete control over every element of their plots, including the colors, line styles, markers, axes, and labels.

Statsmodel : A popular Python library for statistical modeling and analysis is called Statsmodels. It offers a complete collection of tools for carrying out different statistical analyses, such as regression analysis, time series analysis, hypothesis testing, and others.

Users can quickly and simply fit statistical models to their data using Statsmodels to gain insightful knowledge. The library provides a variety of statistical models, including generalized linear models, logistic regression, time series models (such ARIMA and SARIMAX), and many more.

Sklearn : A well-known Python package for machine learning and predictive modeling is called Scikit-learn, or Sklearn. For tasks like classification, regression, clustering, dimensionality reduction, and model selection, it offers a complete set of techniques and tools.

Sklearn provides a uniform and user-friendly API that makes it simple to employ machine learning methods on a variety of datasets. The library has a modular structure, with various modules devoted to various facets of machine learning, such as feature selection, preprocessing, model training, and evaluation.

3.2.2 MAD (Median Absolute Deviation):

The median absolute deviation is a statistical measure of value dispersion. We can easily check how spread out the data is around its median. Standard deviation and variance are also factors of dispersion but they are highly affected by the anomalies in the data which makes our result less reliable.

To apply this method we calculate robust z-scores which can easily be calculated as

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD}$$

formula 4

where, x is the value for which z-score is calculated,

x_i is the current value for which z-score is calculated

$\bar{x}$ is the median of the sample

MAD is the absolute difference between value of the sample and median of the sample

$$MAD = median(|x_i - \tilde{x}|)$$

formula 5

3.2.2.1 Why 0.6745?

In traditional z-score calculation standard deviation is used but in robust z-scores we use median standard deviation which is always smaller than standard deviation so to make the robust z-score look like a traditional z-score we need to scale it.

If we see a normal distribution which has no anomalies, we get MAD as 0.6745. So here we need to divide the MAD with 0.6745. Hence we are multiplying it with the final formula of robust z-scores.

3.2.2.2 How to find anomaly

So we calculate all the z-scores for all the current values and see how each value deviated from the median.

Now we are ready to fix a threshold to take out the anomalies from the data. We use ±3.5 or ±3.0 to flag the anomalies.

3.2.2.3 Why the Threshold is 3 or 3.5

To understand the value of 3 or 3.5 we have to look at the normal distribution or gaussian distribution.

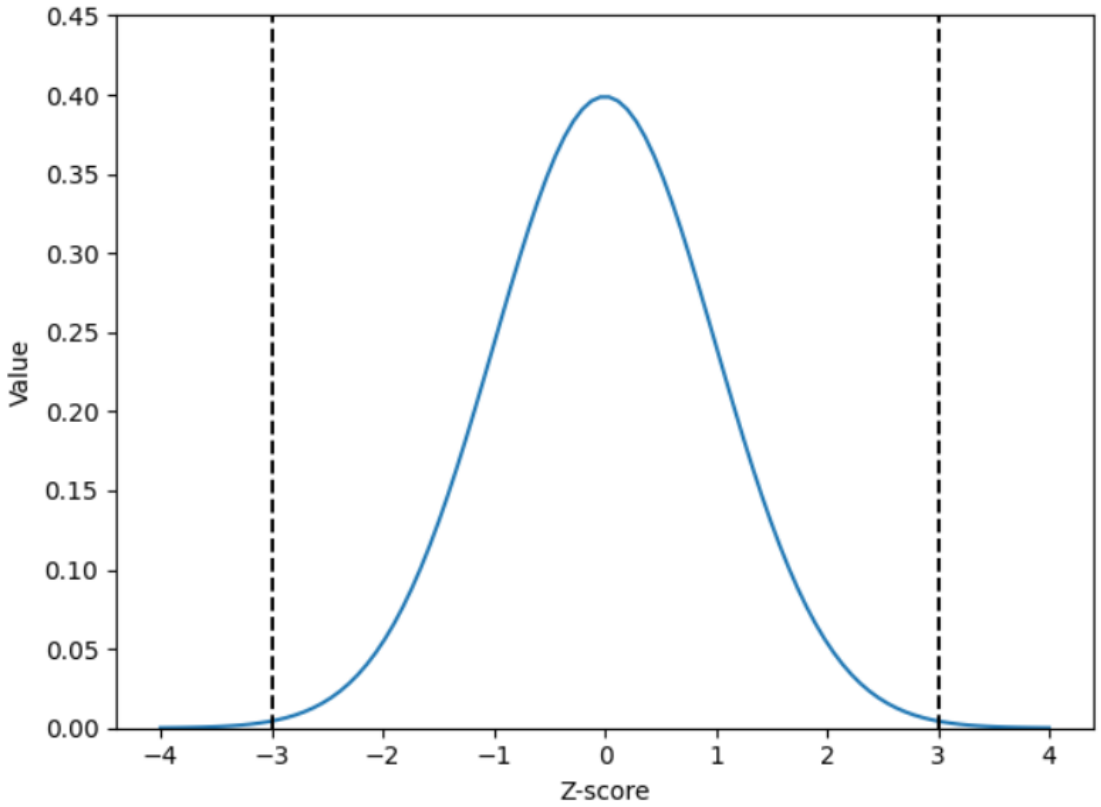

figure 2: Normal distribution

As we see in this Normal distribution the values before -3 and after 3 start to scatter very far away from the mean. So we take ±3.5 or ±3.0 as a threshold for detecting anomalies.

Pros:

1. Very easy to use and resource efficient.
2. does not take anomalies into account while calculating the MAD.

Cons:

1. Works well when we have a perfectly Normal distribution of data.

3.2.2.4 Modules used

Pandas: pandas library is used to handle very big csv files as it gives us all operation's function to manipulate large data in csv files. We have used pandas to create new columns in csv files as well as to peek into the csv files.

Numpy: numpy library is used for complex calculations done in python code. We used numpy to create an array which helped us to plot the normal distribution.

Math: This module helped in taking square root and complex math calculations.

Scipy: This module makes the statistical function very easy to implement. It is used in plotting normal distribution, calculating median absolute deviation and plotting confusion matrix.

Matplotlib: This module was used to plot each and every curve in the implementation.

Seaborn: This module was used to plot the univariate distribution of the data in implementation.

4. Results and Analysis:

4.1 ARIMA MODEL:

We have implemented ARIMA model in different set let me tell one of the dataset which i have used for this ARIMA model that is shampoo_sales which i have taken for ARIMA modeling

Step 1:

We need to initialize different library that is panda , matplotlib , statsmodel , Sklearn and math

So after doing the first step we get the output :

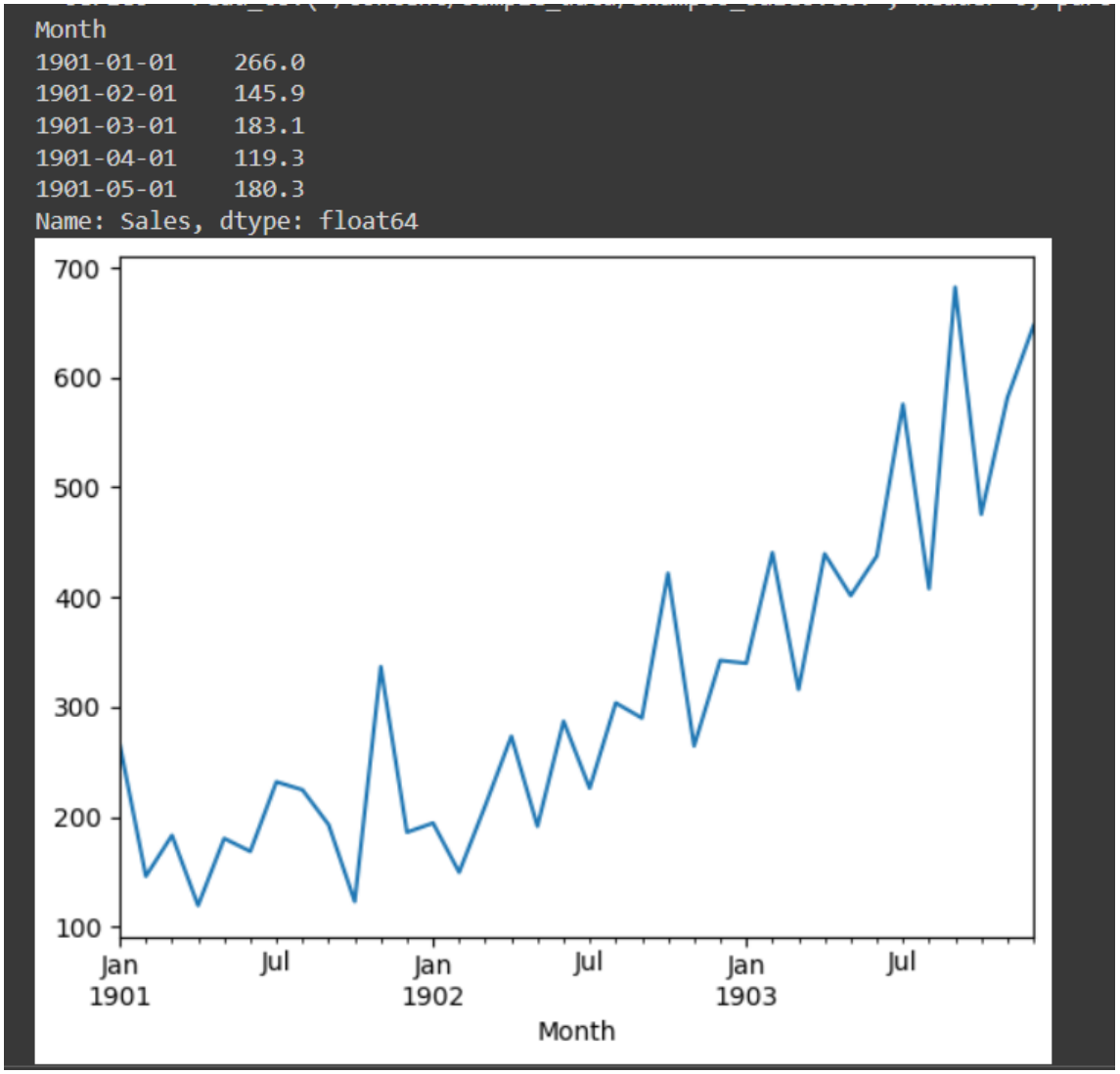

figure 3

This is the raw data which we get after implementing the library and plotting the graph for the dataset

Step 2:

It seems in output that our graph is not stationary so we need to differentiate to make it stationary , it may be a difference of order 1 also .

Let's take a look at autocorrelation plot of time series

The output will come as

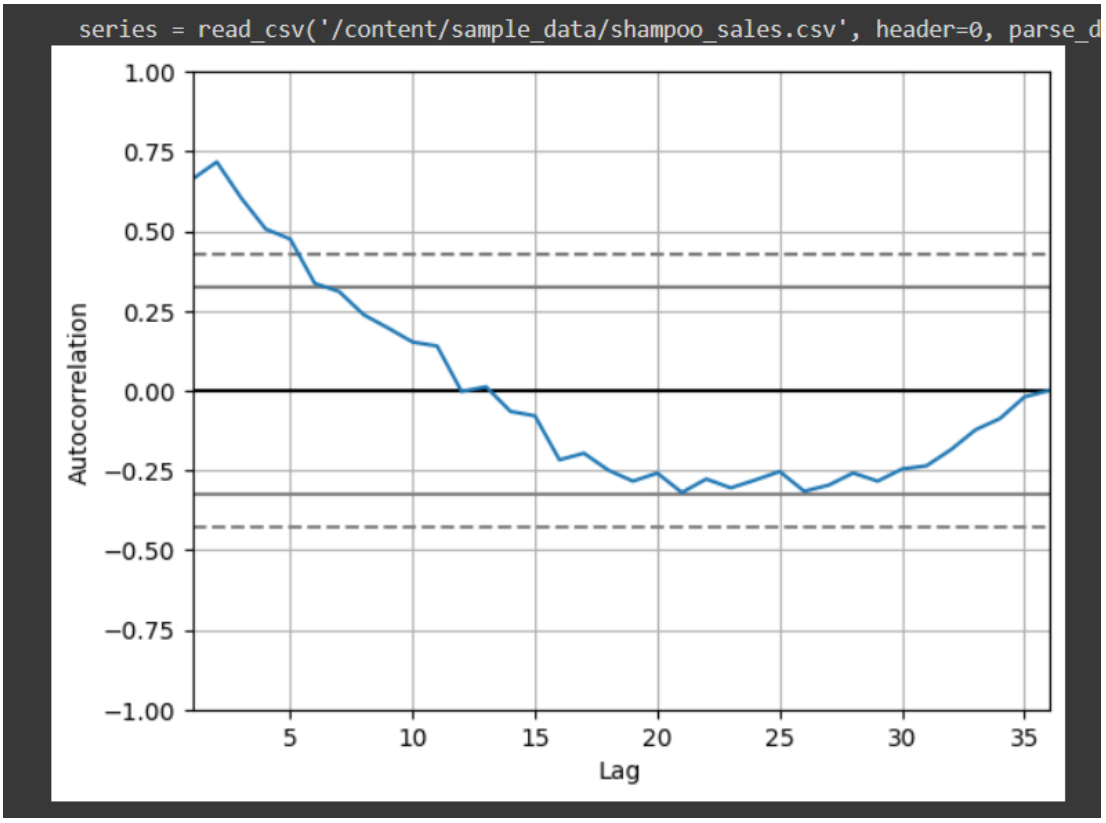

figure 4

Step 3 :

Now after doing autoregression Statsmodel library has module of ARIMA() where we pass the parameter p ,d ,q as for setting value for autoregression , difference order to make it stationary , and moving average

The output will be :

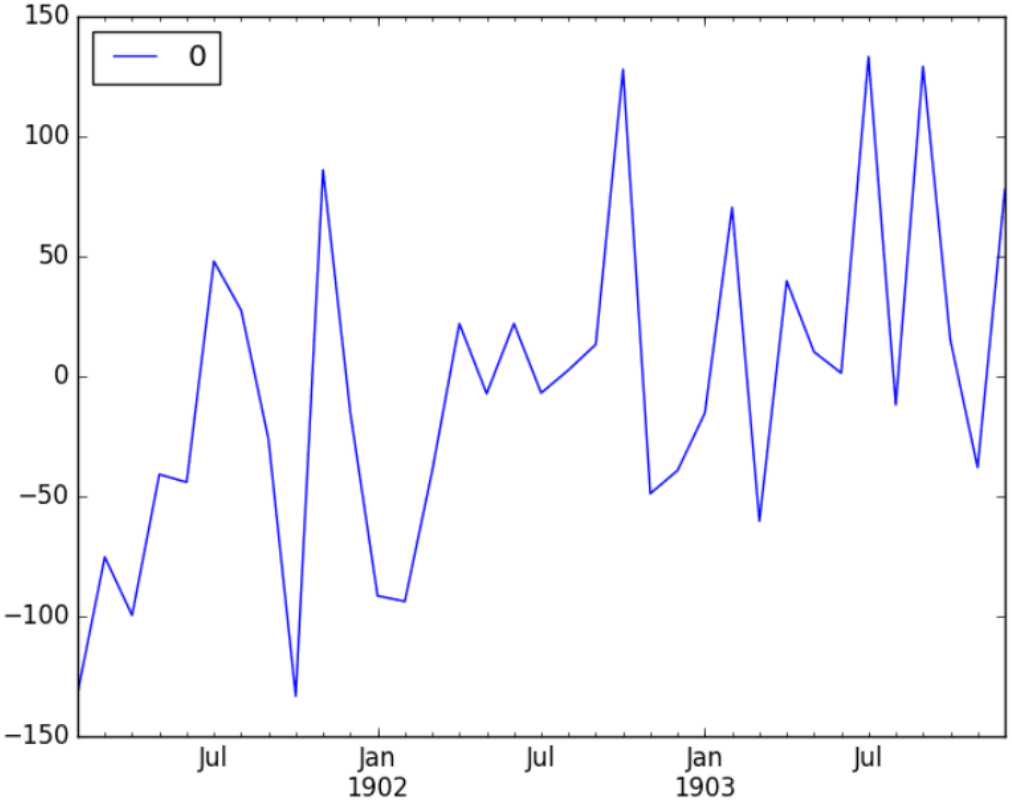

figure 5

And the Density plot of a data is

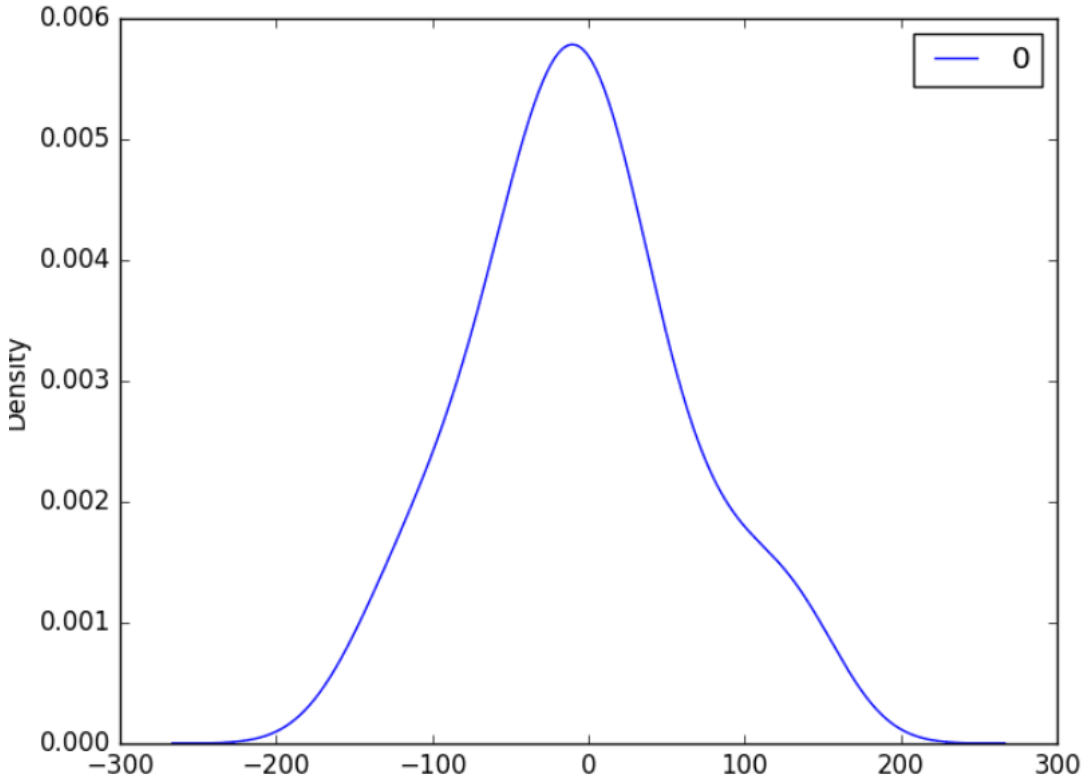

figure 6

Step 4 :

Now we do the forecasting on the dataset

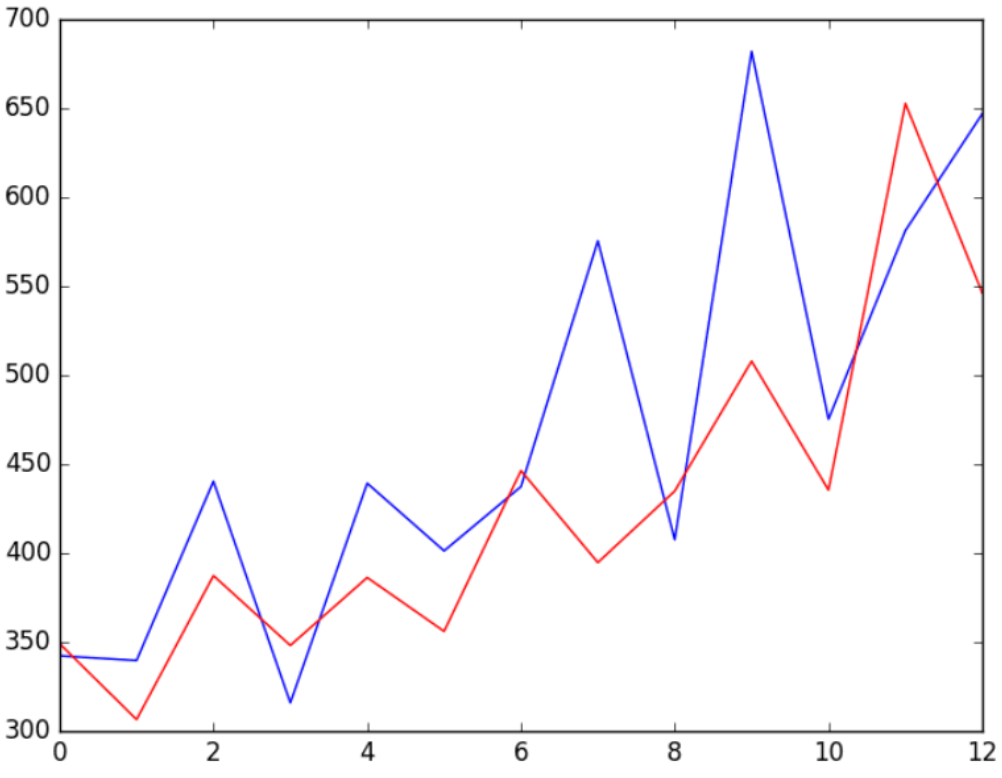

figure 7

4.2 Median absolute deviation:

first we read the dataset and make a new column is_anomaly which is initially assigned value 1.

	timestamp	value	is_anomaly
0	2014-02-14 14:30:00	1.732	1
1	2014-02-14 14:35:00	1.732	1
2	2014-02-14 14:40:00	1.960	1
3	2014-02-14 14:45:00	1.732	1
4	2014-02-14 14:50:00	1.706	1

table 1

Our first five columns of the data look like this.

Now we plot a scatter graph for our data.

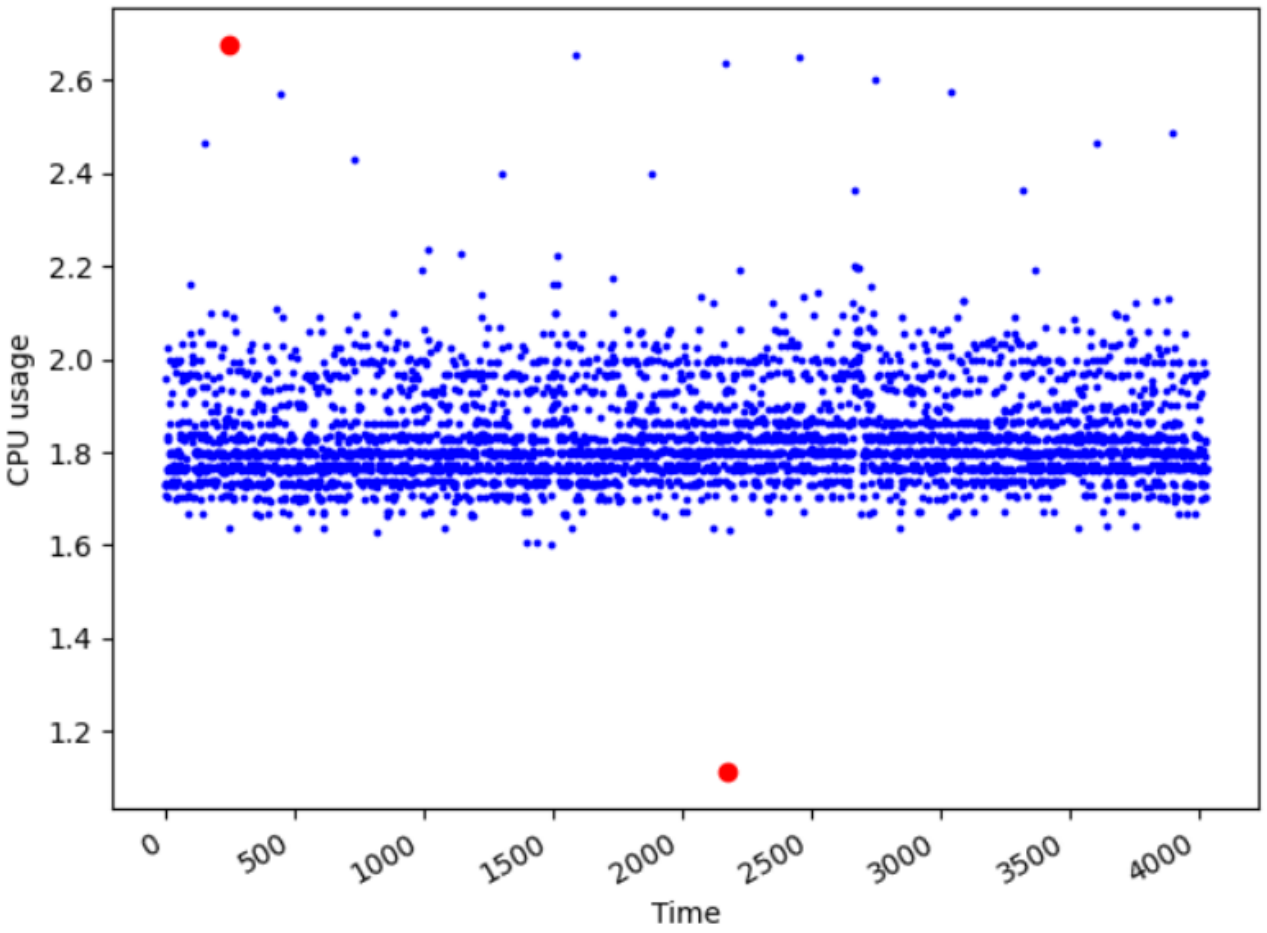

figure 8

In the scatter plot blue dots are the normal data values while the red dots are the anomalies.

With the plot of the distribution function for our dataset we start our robust z-score calculations and then consider the threshold of -3.5 and 3.5, upper and lower threshold respectively.

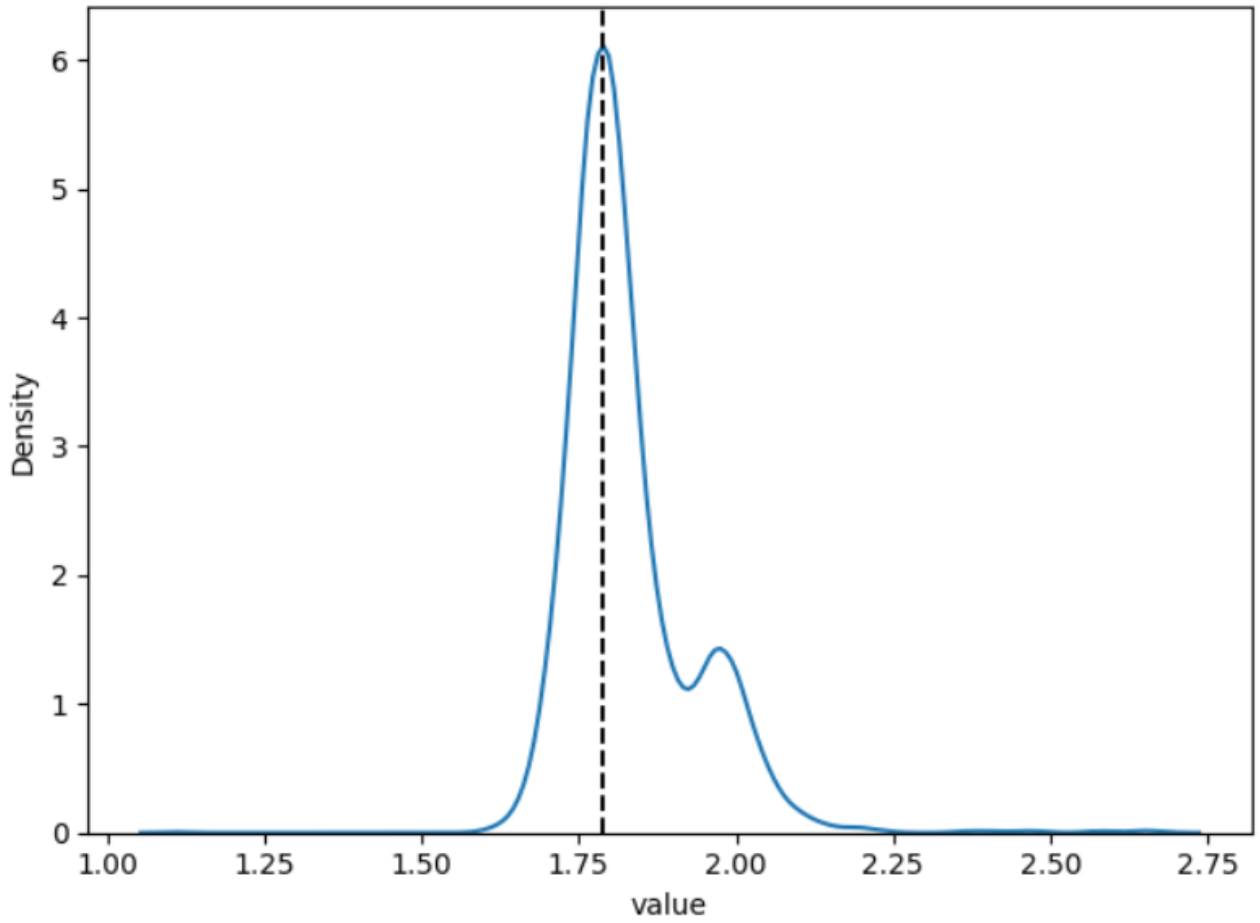

figure 9

	timestamp	value	is_anomaly	z-score
0	2014-02-14 14:30:00	1.732	1	-1.274056
1	2014-02-14 14:35:00	1.732	1	-1.274056
2	2014-02-14 14:40:00	1.960	1	2.997778
3	2014-02-14 14:45:00	1.732	1	-1.274056
4	2014-02-14 14:50:00	1.706	1	-1.761194

table 2

Now with the comparison of the z-score values we start the anomaly detection which we will present in the form of a confusion matrix.

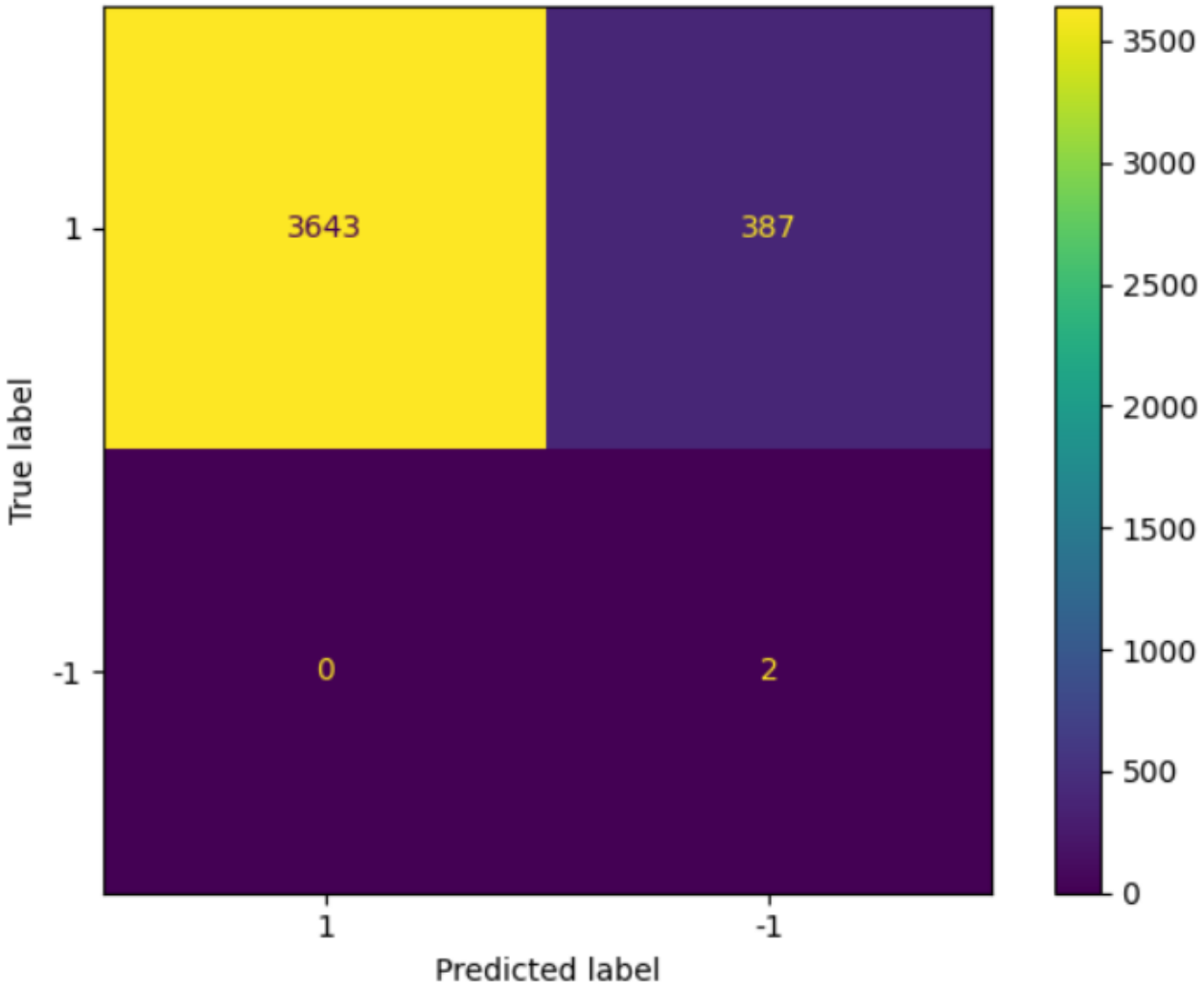

figure 10

this resultant confusion matrix represent 4 sections:

1.predicted label: 1 and true label: 1: This section indicates that the data points which are not actually anomalies are predicted also as non anomalies. The number of such points is 3643.

2. predicted label: 1 and true label: -1: This section represents the data points which are actually predicted as not anomalies, also known as false negatives. False negatives are zero in this case.

3.predicted label:-1 and true label: -1: This area represents the true outcomes of the method, in this area are the points which are anomalies and predicted also as anomalies. In reality there were two anomalies in the data and both are predicted correctly.

4. predicted label:-1 and true label:1: This area represents false positives,i,e the points which are not anomalies but are predicted as anomalies. Total 387 false positives are there in the data.

5. Conclusion and Future Scope:

Time series detection is a tedious task and indeed it takes time and patience to find or develop a new algorithm, but if we look very closely we will observe that in the time of Machine learning why do we use forecasting or statistical methods? The answer to this question lies in the fact that the resources are scarce and not everyone holds the processing power to carry out the Machine learning algorithms. From the above results we can see that these old school methods are not outdated but they need some grooming to make them better. It is very clear that they cannot surpass advanced machine learning algorithms but can be optimal alternatives to those who cannot afford the resources.

The MAD model is the model to go for when the anomalies are large as it is clear from the result that it does not take the whole data into account while finding anomalies. So it is better to use in case of more anomalies whereas ARIMA model is best to use in the case when the anomalies are lower in number so that the average is least affected with the anomaly ridden data points and a perfect forecast can be obtained. Both of these algorithms are not perfect but require a little more work, as in case of MAD model it requires a method to develop two different bounds for the baseline selection in the case when the data is not normally distributed. ARIMA model can be combined with other different models to predict a value very close to the actual value.

References

[1704.07706] Automatic Anomaly Detection in the Cloud Via Statistical Learning. (2017, April 24). arXiv. Retrieved May 29, 2023, from https://arxiv.org/abs/1704.07706

Anomaly detection on social media using ARIMA models. (n.d.). DiVA portal. Retrieved May 29, 2023, from http://www.diva-portal.org/smash/get/diva2:882392/FULLTEXT01.pdf

Braei, M., & Wagner, S. (2020, April 1). *[2004.00433] Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art*. arXiv. Retrieved May 29, 2023, from https://arxiv.org/abs/2004.00433

Kozitsin, V., Katser, I., & Lakontsev, D. (2021, April 2). *Online Forecasting and Anomaly Detection Based on the ARIMA Model*. MDPI. Retrieved May 29, 2023, from https://www.mdpi.com/2076-3417/11/7/3194

Park, H. (n.d.). *Analysis of Anomaly Detection Methods for Streaming Data*. Open Access LMU. Retrieved May 29, 2023, from https://epub.ub.uni-muenchen.de/68482/1/BA_Park_Hyeyoung.pdf