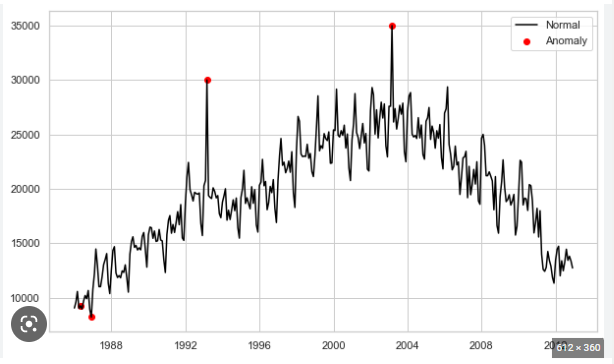
**Anomaly Detection**

**Anomaly:** It is different or abnormal and deviated substantially from other data in the sample/historical data.

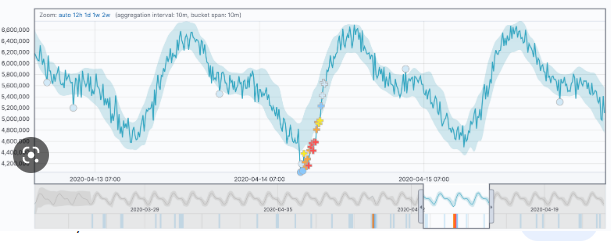
1. **Point Anomalies:** Single instances of something abnormal

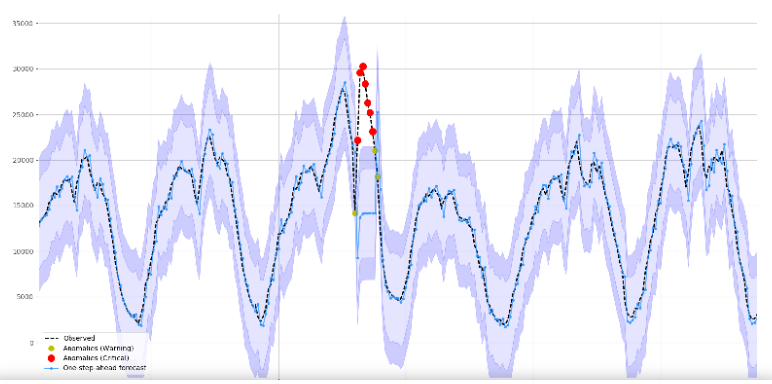
* Eg: in a stock price time series, a point anomaly could be a sudden and sharp increase or decrease in stock price for a particular day or period, which does not follow the usual trend or pattern of the stock price over time
* various applications, such as fraud detection, **fault diagnosis**, and anomaly detection in industrial processes.

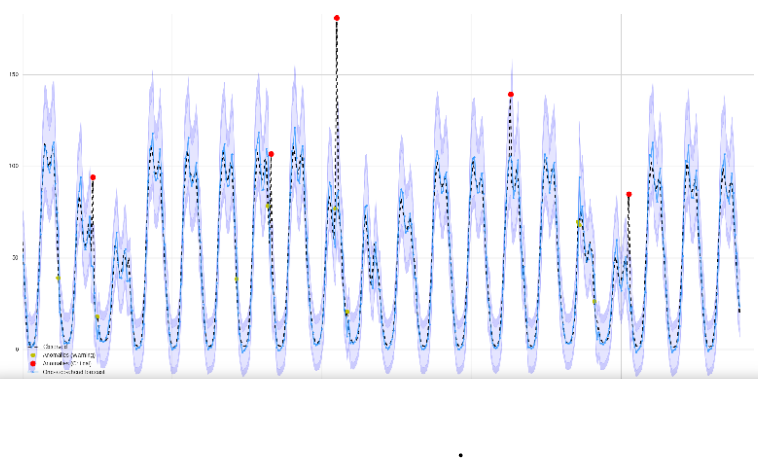


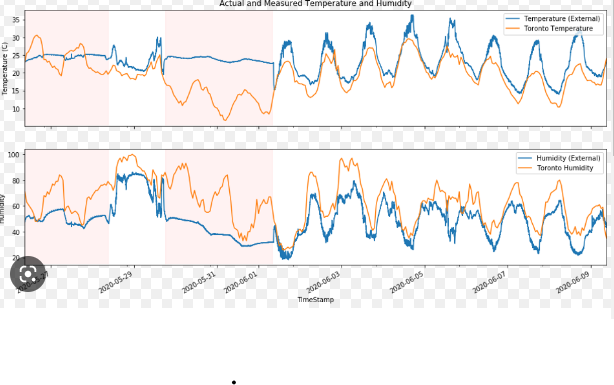
1. **Pattern Anomalies:** Clusters of data that are abnormal. A pattern anomaly in time series data occurs when there is a significant deviation in the regular pattern or trend of the data.

* In other words, it refers to a deviation from the expected shape or structure of the time series.
* Eg: **Industrial process monitoring:** Pattern anomaly detection can be used to detect anomalies in the patterns of sensors and machines in industrial processes. This can help in identifying potential faults or malfunctions in equipment and prevent downtime.







****

**Time Series Anomaly detection:**

* Time series anomaly detection is the process of identifying patterns in time series data that deviate significantly from the expected or normal behavior.
* Anomalies are data points or patterns that are rare, unexpected, or do not conform to the regular pattern or trend of the time series.
* This process is useful in many applications such as fraud detection, intrusion detection, equipment failure prediction, and medical diagnosis.
* The goal is to detect anomalous behavior as early as possible so that appropriate action can be taken to mitigate or prevent negative consequences.
* **Local Outlier:** A local outlier is an observation that deviates significantly from its neighboring observations in the feature space. Usually common.
* **Global outlier:** on the other hand, is an observation that deviates significantly from the majority of the observations in the dataset. Global outliers are typically detected using statistical methods such as Z-score and model-based methods such as One-Class SVM (OCSVM), Isolation Forest, Pycaret and Minimum Covariance Determinant (MCD).
  + For example, in a dataset of blood pressure measurements, a global outlier may represent a patient with an extremely high or low blood pressure reading.

**Anomaly data points & Novelty data points:**

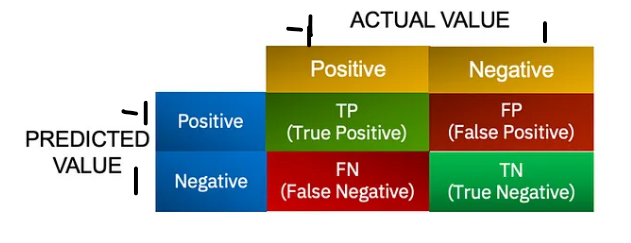
1. **Anomaly Data points:** Anomaly data points are data points that deviate from the norm in a given dataset. These can be considered local outliers because they are anomalous within a particular subset of the data. For example, in a dataset of daily temperatures, an anomalous data point might be a particularly hot or cold day that is significantly different from the other days in the same month.
2. **Novelty data points:** Data points that are significantly different from the majority of the data in a dataset as a whole. These can be considered global outliers because they are novel in the entire dataset, rather than just within a particular subset. For example, in a dataset of daily temperatures, a novelty data point might be a day in which the temperature is significantly higher or lower than any temperature that has ever been recorded in that location.

**List of Algorithms can be used:**

1. **Unsupervised and model based:** These models are built explicitly to find the anomalies. Eg: Isolation Forest, One-class SVM, Pycaret etc
2. **Statistical based Methods:** Statistical methods are based on the assumption that normal data follows a specific statistical distribution. Any data point that deviates significantly from the expected statistical distribution is considered an anomaly. Examples of statistical methods include Z-score, Boxplot.
3. **Deep Learning-based Methods:** Deep learning-based methods, such as Autoencoder, use neural networks to learn the patterns in the data and detect anomalies.

**Confusion Matrices:**

**‘-1’ : Anomaly, ‘1’ : Normal point (not anomaly)**



TP: Number of Anomalous point identified as Anomaly

TN: Number of Normal points identified as normal point

**FP: Number of Normal points identified as Anomaly**

**FN: Number of Anomalous point identified as Normal point**

Using this confusion matrix, we can calculate various performance metrics to monitor the machine's performance, such as:

1. Accuracy: The percentage of correct predictions.
   * + Accuracy = (TP + TN) / (TP + FP + TN + FN)
2. Precision: The percentage of correct anomaly predictions out of all predicted anomalies.
   * + Precision = TP / (TP + FP)
3. **Recall or Sensitivity:** The percentage of correctly identified anomalies out of all actual anomalies.
   * + Recall = TP / (TP + FN)
4. F1 Score: The harmonic mean of precision and recall.
   * + F1 Score = 2 \* (Precision \* Recall) / (Precision + Recall)

**Pre-requisites:**

* Dataset – Artificially generated data based on 5-Second frequency
* Dataset – Create spikes

**Pipeline:** Anomaly detection is the process of identifying patterns in data that do not conform to expected behavior. There are many algorithms that can be used for anomaly detection, but one popular approach is to use machine learning algorithms to learn the patterns of normal behavior and then use these patterns to identify anomalies. Here is a general algorithm for anomaly detection:

1. **Collect data**: Gather data on the system or process that you want to monitor for anomalies.
2. **Preprocess data:** Clean the data and transform it into a format that is suitable for analysis. This may involve removing outliers, normalizing the data, or reducing the dimensionality of the data.
3. **Split data:** Split the data into training and test sets. The training set will be used to train the anomaly detection algorithm, while the test set will be used to evaluate the performance of the algorithm.
4. **Train algorithm**: Train a machine learning algorithm on the training set. Some popular algorithms for anomaly detection include clustering algorithms, support vector machines, and deep learning algorithms.
5. **Evaluate performance:** Evaluate the performance of the algorithm on the test set. This may involve calculating metrics such as precision, recall, and F1 score.
6. **Tune algorithm:** Adjust the parameters of the algorithm to optimize its performance. This may involve adjusting the threshold for what constitutes an anomaly or experimenting with different algorithms.
7. **Deploy algorithm:** Deploy the anomaly detection algorithm in the system or process that you want to monitor for anomalies. This may involve setting up automated alerts or integrating the algorithm into a larger monitoring system.
8. **Monitor system**: Continuously monitor the system or process for anomalies using the deployed algorithm. If an anomaly is detected, investigate the cause and take appropriate action.

Top of Form

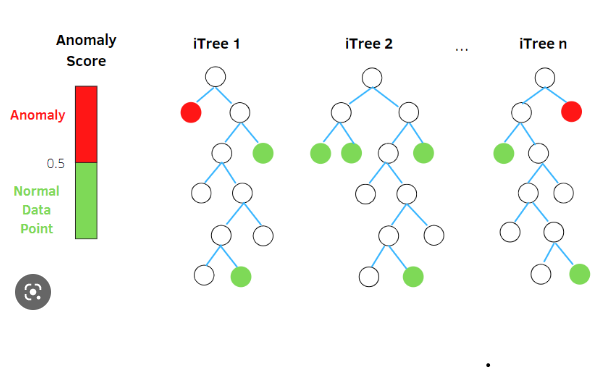
Bottom of Form

1. **Isolation Forest Anomaly Detecting Algorithm**

It is explicitly designed to find the anomalies. It is unsupervised algorithm that uses a decision tree to isolate anomalous data points. It works by randomly partitioning the feature space and using the number of partitions required to isolate a data points as an indicator of its anomalousness. Anomaly point require fewer splits from rest.

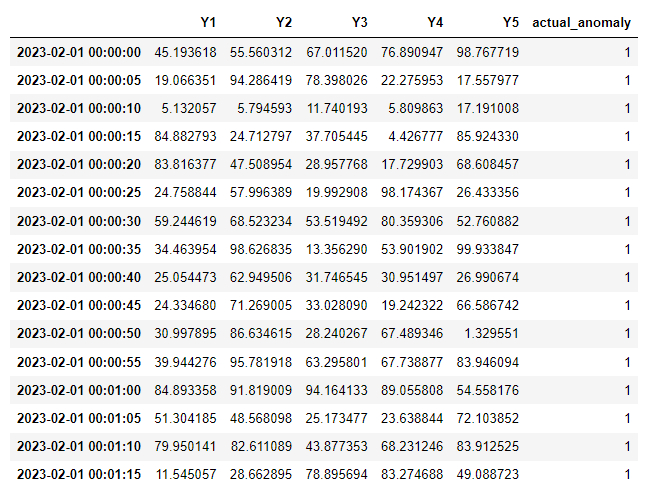
Anomalies isolation is implemented without employing any distance or density measure.

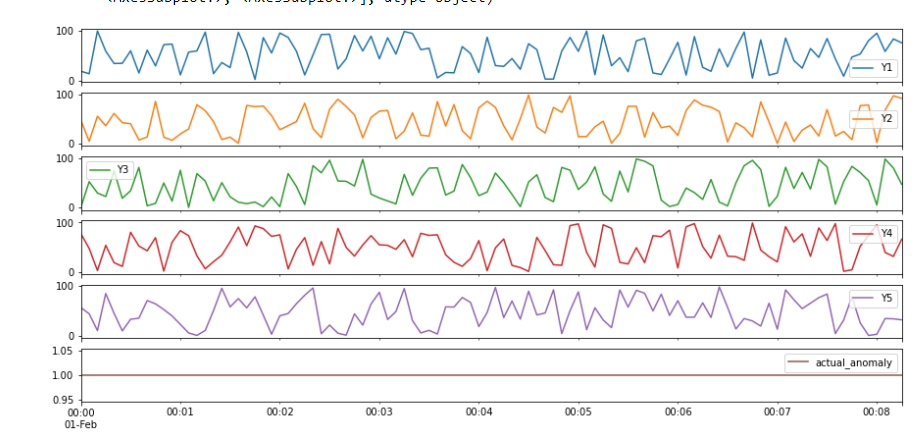
* We set contamination=outlier\_fraction of outliers are present in data. It is trial/error metric
* Perform outlier detection. It returns ‘1’ for normal point, ‘-1’ for anomaly point.

****

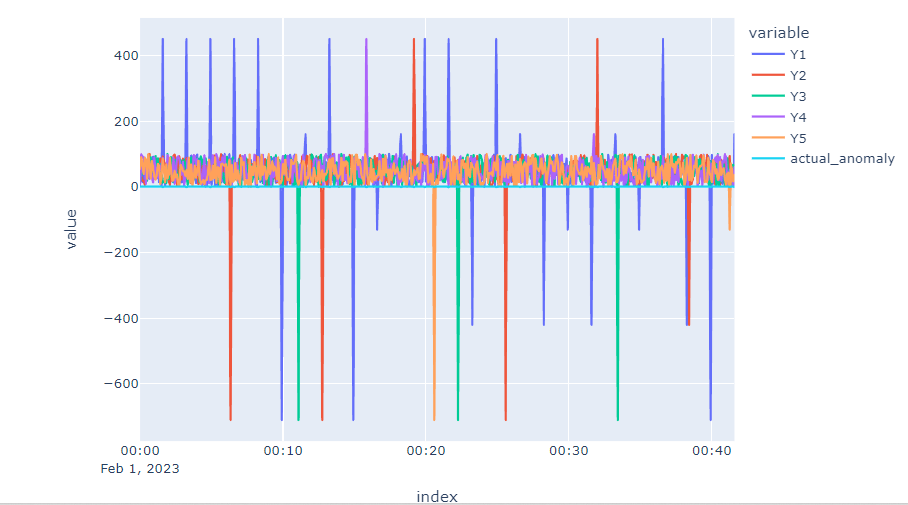
**Code snapshots:**

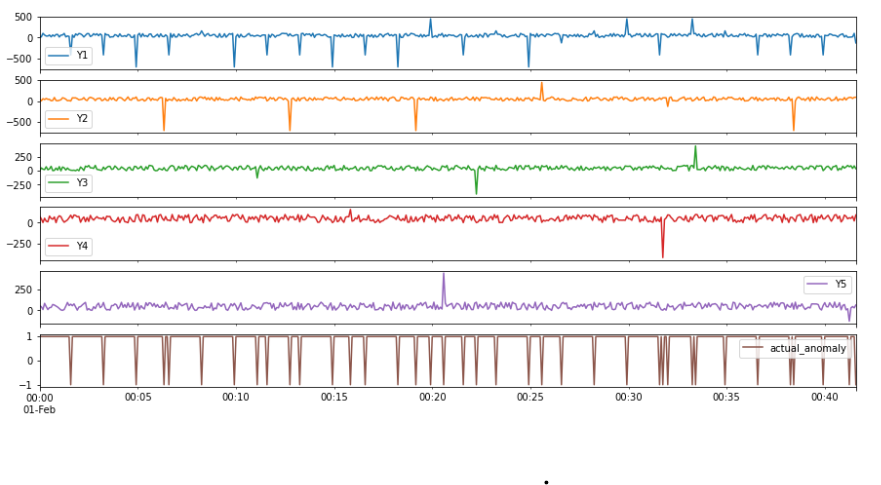
Artificially Data Generated at frequency=5 Seconds for 100 hours. (Total Records= 69000)



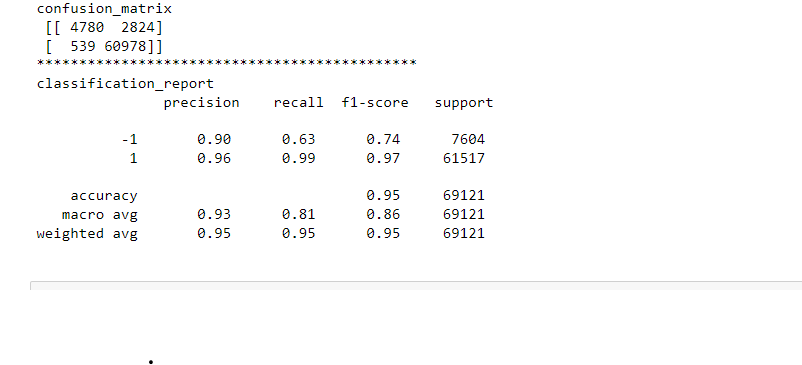


Created spikes/drops at particular frequencies for each features:

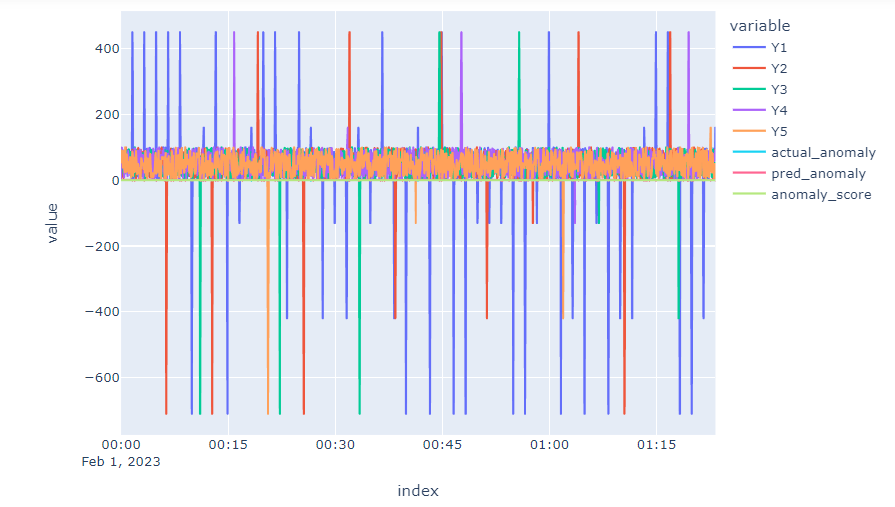


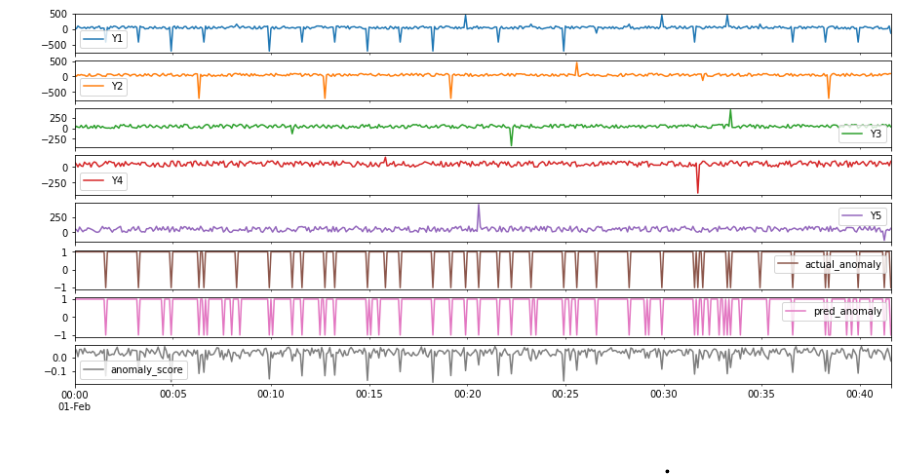


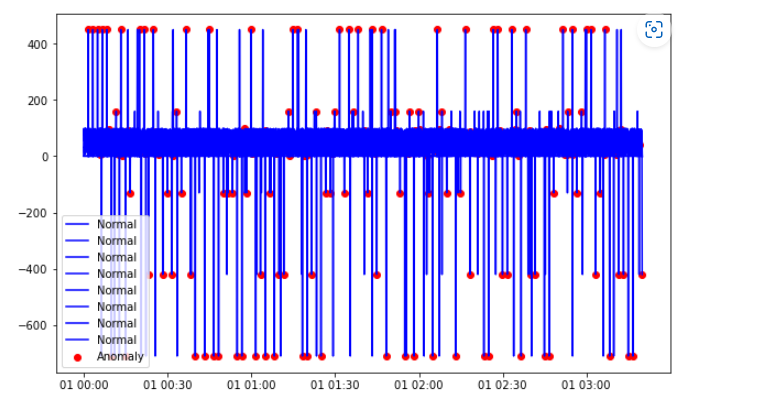
Accuracy of prediction:



Anomaly visualization:





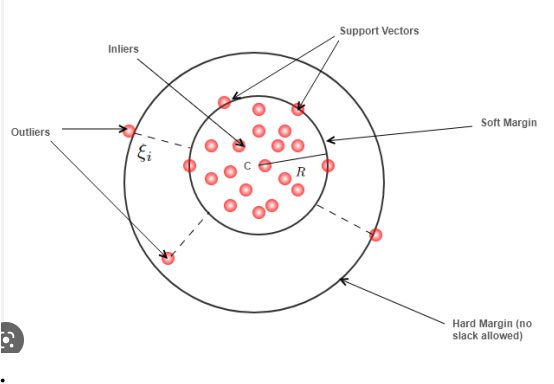


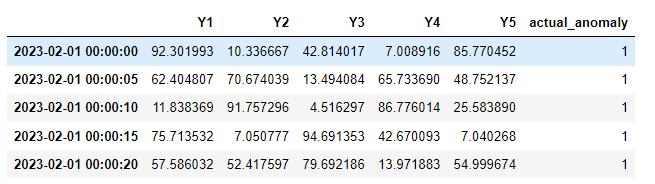
1. **One Class SVM Anomaly detection Algorithm:**

* This algorithm is trained on a dataset that only contains normal data points and it uses this information to identify anomalies in new data points. One-Class SVM (Support Vector Machine) is often used for pattern-based anomaly detection. This is because One-Class SVM is a type of unsupervised learning algorithm that can identify anomalies in a dataset without being explicitly trained on what constitutes normal or anomalous behavior.
* In pattern-based anomaly detection, the algorithm is trained on a set of features that represent normal behavior or patterns in a dataset. Then, when presented with new data, the algorithm can determine whether it fits within the expected patterns or deviates significantly from them.
* One-Class SVM is particularly useful for this type of anomaly detection because it works well with high-dimensional datasets and can identify non-linear patterns in the data.
* Main advantage: train the classifier using only patterns belonging to the target class distribution.

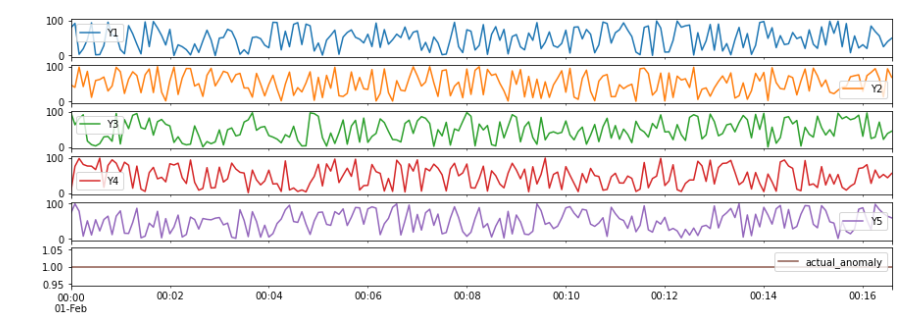
**Parameters to focus on:**

* **Nu**=[0 to 1] controls the fraction of training data that is considered to be outliers.
* **Decision threshold**: identify decision threshold using decision\_function() method. Data points with the score lower than decision threshold are considered novel.
* **Kernel** = ‘rbf’ Radial basis function is used for non-linear problems

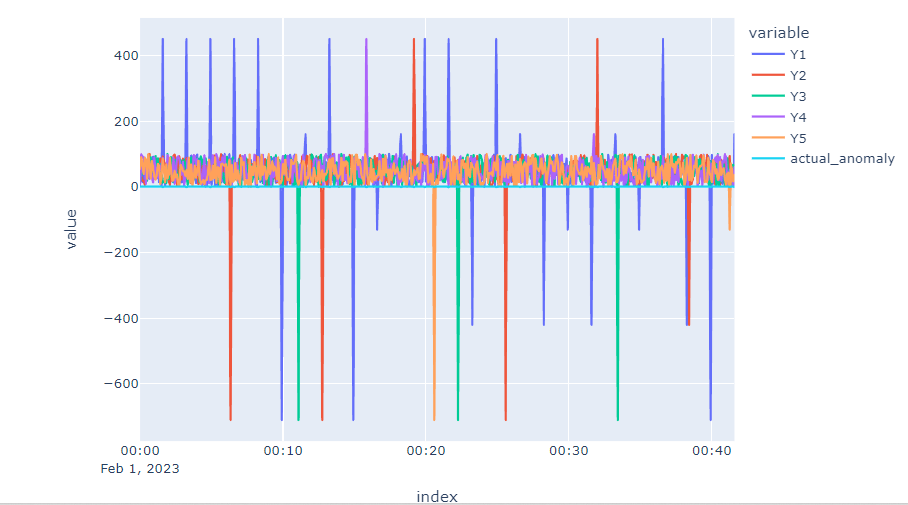


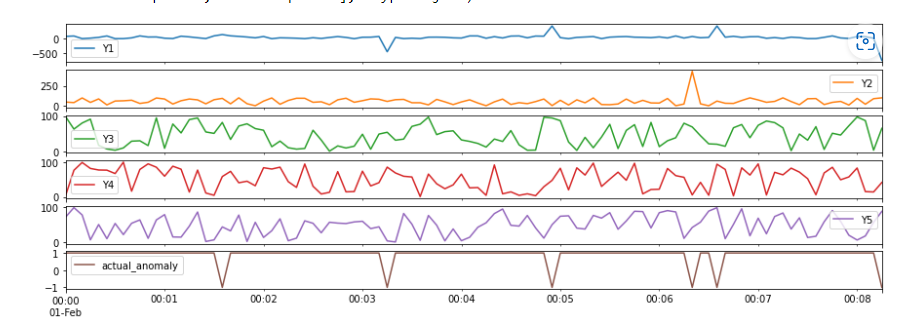
Code snap shots: 

Artificially Data Generated at frequency=5 Seconds for 100 hours. (Total Records= 69000)

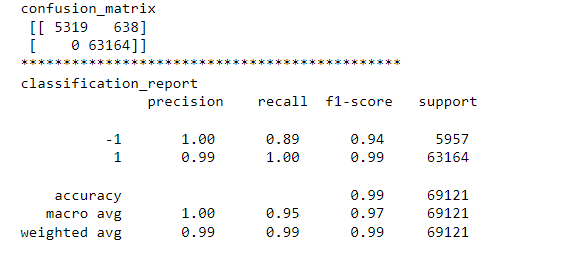


Created spikes/drops at particular frequencies for each features:





Accuracy of prediction:



Anomaly visualization:

