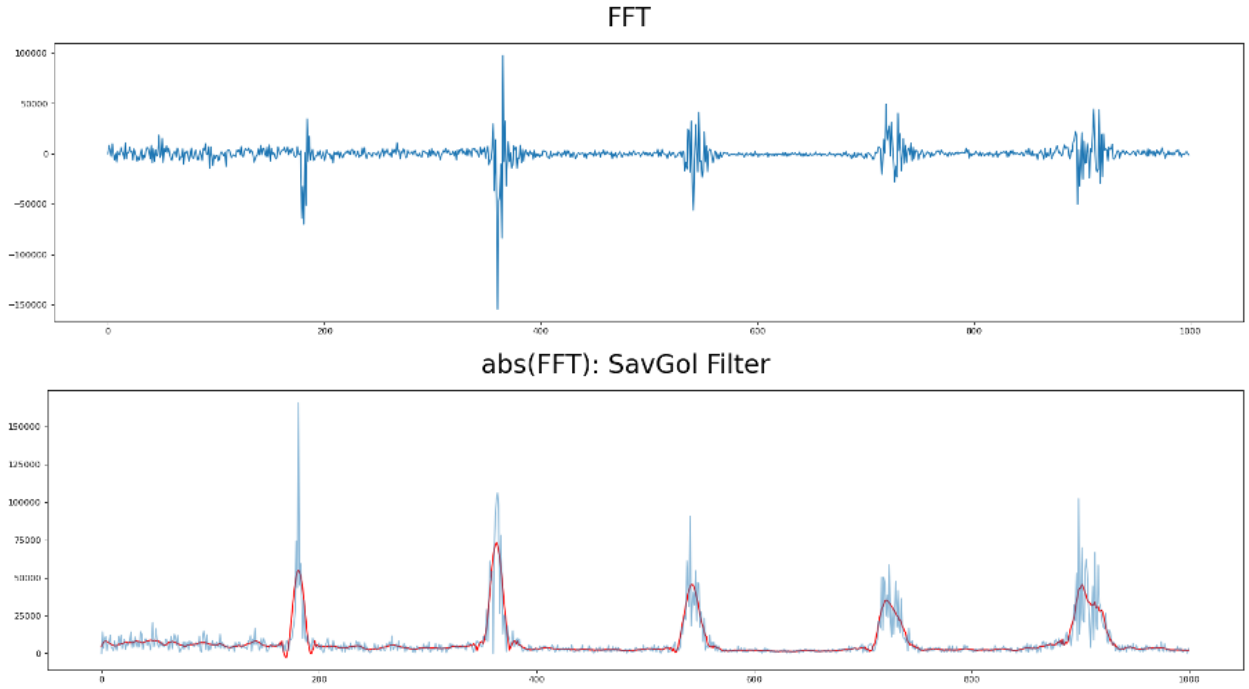
Time Series Anomaly Detection Project

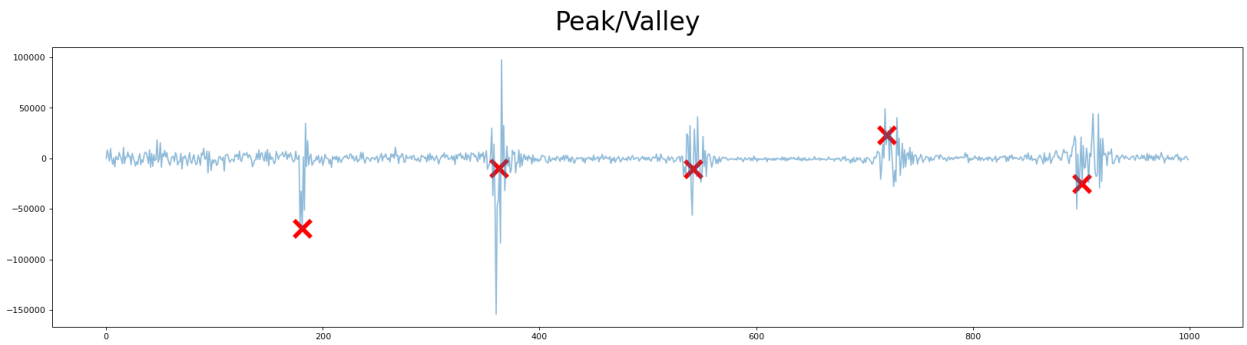
CHOI Jang Hyeon, 20783567, jhchoiac@connect.ust.hk

Data Engineering Method

When handling a time series data, there is a periodicity pattern. The dataset may be a year-long data with seasonal trends, or it could be a daily record of temperature. It is thus important to find the period of the data to facilitate anomaly detection by algorithms. One of the methods of detecting periodicity is via Fourier Transform [4].

The variance detection scoring algorithm is best utilized when provided with a period, and this period is calculated by applying the Fourier Transform on the dataset to find the optimum window size. The period of the dataset would be the first peak in the frequency domain. However, the Fourier transform of the dataset includes numerous peaks, so a Savitzky-Golay filter is used for smoothening after taking the absolute value of the Fourier transform; the absolute value is taken to consider not only peaks but also valleys for convenience.





(Peak locations are indicated by red crosses. First peak at ~180.)

(Estimated period of 180 in the dataset snippet denoted by a straight line.)

The period can thus be estimated using the Fourier Transform method, and the rolling variance produces a graph that accentuates the anomaly in the testing portion of the data. It was noticed, however, that periods shorter than 100 were difficult to detect.

Model

Libraries

The following are a list of libraries that were used for each scoring algorithm.

Rolling min-max	**Z-score**	**Variance detection**	**Matrix Profile**
numpy	scipy.stats.zscore	scipy.stats	stumpy.stump
		numpy.fft	
		scipy.signal.find_peaks	
		scipy.signal.savgol_filter	
		pandas.rolling.var()	

[REDACTED]

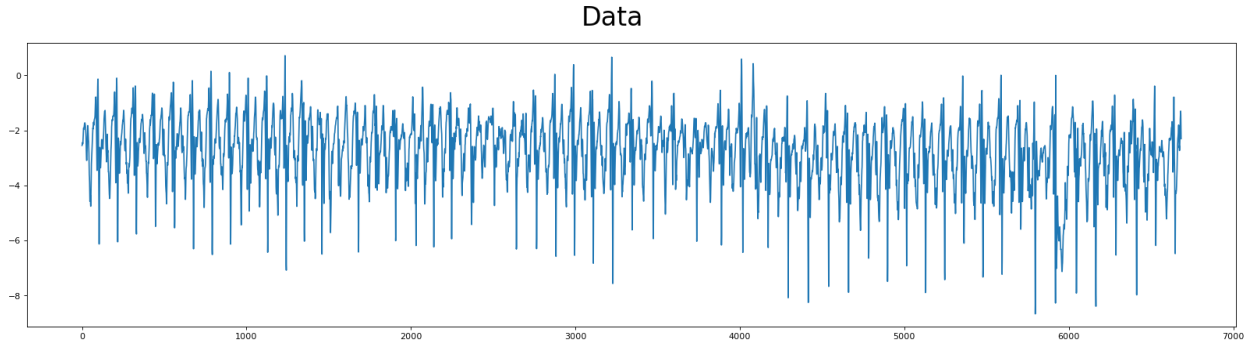
Z-score

The Z-score of the dataset is calculated to determine the outlier's relative value against the rest and it is known to be an efficient method to detect outliers [1]. This scoring algorithm mainly focuses on point anomalies in the dataset by measuring the standard deviation of each value from the dataset mean. The higher the standard deviation from the mean, the more likely it is an outlier. In this project, the absolute value of the z-score has been taken for easier comparison against each other.

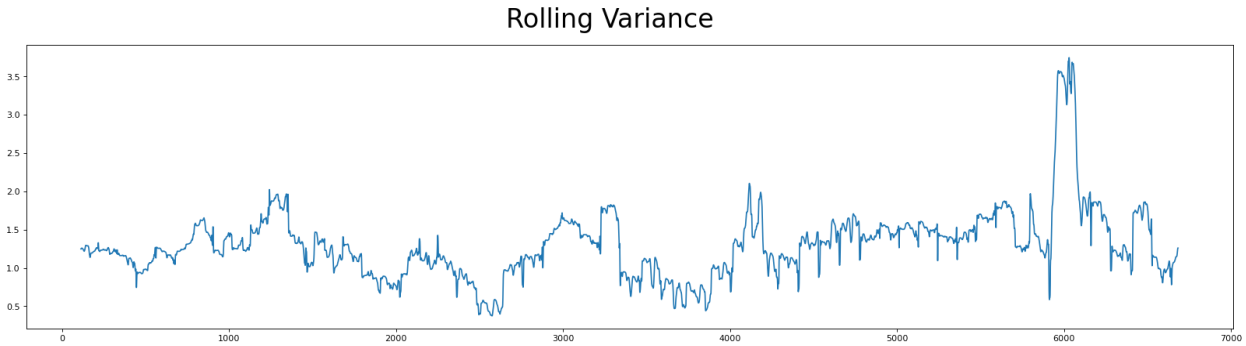
Variance detection

A rolling variance method is used for this scoring algorithm. Rolling variance calculates the variance within a certain window range throughout the entire data. The window range is derived by preprocessing the data into the frequency domain using Fourier transform.
After storing all the variances one-by-one into a list, the z-score of this list is calculated to find the index with the highest variance.

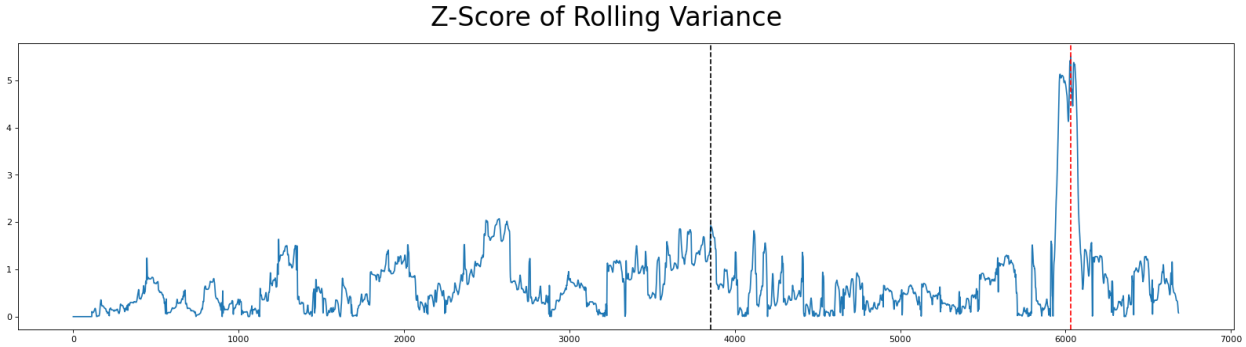
The following graphs illustrate how the variance detection scoring algorithm works. After retrieving a window size from the Fourier Transform, a rolling variance is performed on the dataset. Intuitively, the varying section of the dataset will be detected by the rolling variance algorithm if the period is correct. The algorithm then converts the rolling variance values into z-scores (notice how the y-axis changes). Since the highest maximum and the second highest maximum values are quite different, the algorithm will output a high confidence score for the predicted anomaly location of approximately 6000, which actually corresponds to the abnormal region in the testing dataset.



(Entire dataset.)



(Rolling Variance of the dataset with a window size of ~110.)



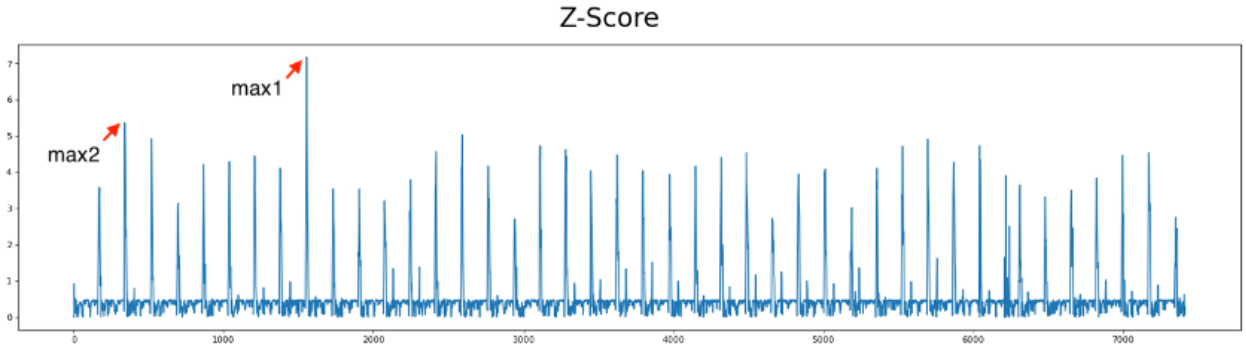
(Z-score of rolling variance. Highest max is indicated by the dashed red line, and the second highest max is indicated by the dashed black line.)

[REDACTED]

Model's Performance

Ensemble of Scoring Algorithms

For the best model performance, the confidence of each scoring algorithm is compared such that the program outputs the most likely anomaly position. Each scoring algorithm measures its confidence level, which is calculated in one of two ways. The first method takes the highest anomaly score and subtracts it by the second highest anomaly score. It is then divided by the first highest anomaly score to normalize it between 0 and 1. In other words, the more different the two anomaly scores are, the more likely it is an outlier; an anomaly score is simply the value of the scoring system, such as variance, or z-score.



(Z-score graph with its first and second highest values annotated.)

$$confidence = \frac{max1 - max2}{max1}$$

(Modified confidence formula based on [2])

[REDACTED]

References

1. Chikodili N.B., Abdulmalik M.D., Abisoye O.A., Bashir S.A. (2021) Outlier Detection in Multivariate Time Series Data Using a Fusion of K-Medoid, Standardized Euclidean Distance and Z-Score. In: Misra S., Muhammad-Bello B. (eds) Information and Communication Technology and Applications. ICTA 2020. Communications in Computer and Information Science, vol 1350. Springer, Cham. https://doi.org/10.1007/978-3-030-69143-1_21

2. KDD Cup 2021 MDTS competition - Team Old Captain. (2021). Retrieved November 15, 2021, from https://www.youtube.com/watch?v=4PdlUcmwWu0.

3. Yeh, Chin-Chia Michael & Zhu, Yan & Ulanova, Liudmila & Begum, Nurjahan & Ding, Yifei & Dau, Anh & Silva, Diego & Mueen, Abdullah & Keogh, Eamonn. (2016). Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. 1317-1322. 10.1109/ICDM.2016.0179.

4. Yousif, Muneera & Celik, Mete. (2019). Developing Fast Techniques for Periodicity Analysis of Time Series. 1-5. 10.1109/ISMSIT.2019.8932865.

5. Zivot, E. and Wang, J., 2006. Modeling Financial Time Series with S-PLUS. New York, NY: Springer New York pp.313-314

6. Zhu, Y., Zimmerman, Z., Shakibay Senobari, N., Yeh, C., Funning, G., Mueen, A., Brisk, P. and Keogh, E., 2017. Exploiting a novel algorithm and GPUs to break the ten quadrillion pairwise comparisons barrier for time series motifs and joins. Knowledge and Information Systems, 54(1), pp.203-236.