Identify Anomalous Behaviour by Analyzing Web Access Logs

Anomaly Detection in Time Series

Thenuka Thanabalasingam
Owen Wang
Runhe Zhong

Contents

1. Background
2. Project objectives
3. One-Class SVM & Isolation Forest
4. Challenges during execution
5. Results & Discussion
6. Lesson learnt

Background

Background

Monitoring server logs could identify server outage or possible cyber attacks at an early stage.

According to Comparitech, 78% of Canadian companies were under cyber attack at least once in 2020. In 2021, the number of Canadian companies experienced cyber attack has increased to 85.7% (A. O'Driscoll, August 1st, 2022).

In the third quarter of 2022, ransom DDoS attacks has increased by 67% compare with the same time last year and with an increase of 24% if compare with the previous quarter. More than 20% of the DDoS attacks fall into the multi-vector DDoS attack category (S. Cook, November 6th, 2022).

Project objectives

Project objectives

To build a machine learning model to identify abnormal activities in web access logs

One-Class SVM & Isolation Forest

One-Class SVM

One-Class SVM is similar to basic SVMs, but it only has one class. A boundary is set based on current data. It detects an outlier, if there is any, when new data comes in.

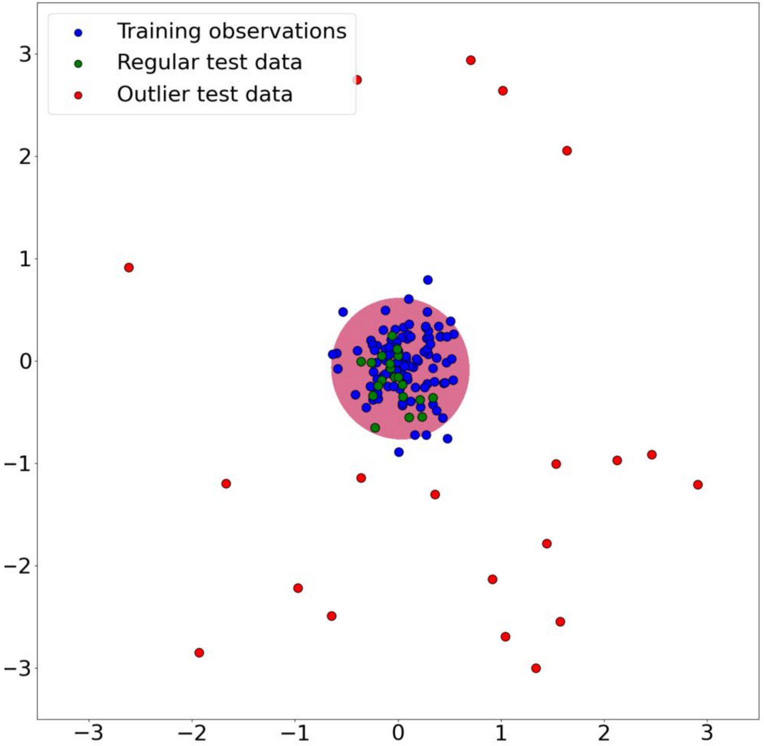

Figure 1. Simple Example Results. from *What is One Class SVM and How Does It Work?* (https://www.baeldung.com/cs/one-class-svm)

Isolation Forest

Isolation Forest is similar to Random Forest. They are all based on decision trees. It randomly selects data and process them with random features. In each tree, the branches that are shorter are (potential) anomalies, because it is easier to isolate them.

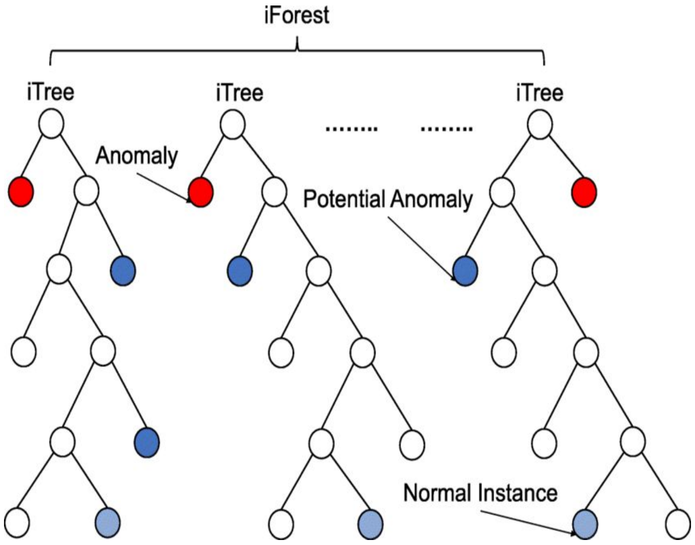

Figure 2. Regaya, Yousra & Fadli, Fodil & Amira, Abbes. (2021). Point-Denoise: Unsupervised outlier detection for 3D point clouds enhancement. Multimedia Tools and Applications. 80. 1-17. 10.1007/s11042-021-10924-x.

Challenges during execution

How to confirm if our algorithms work -- NYC Taxi Traffic

Challenge: **unable to identify anomalies** directly from the data for evaluation

Solution: first build an ML anomaly detection strategy on a dataset **where we already know the anomalies** - NYC Taxi Traffic dataset & then apply it to ECLog Data

- Smaller dataset
- Validated anomalies part of dataset description (on Christmas, historic snow storm & Labour Day) that can be annotated

Additionally, this will make our model **generic** and **easy to translate to other datasets**.

Results & Discussion

NYC Taxi Dataset
-- One-Class SVM

Please refer to our code

NYC Taxi Dataset -- Isolation Forest

Please refer to our code

E-Commerce server access logs -- One-Class SVM

Please refer to our code

E-Commerce server access logs -- Isolation Forest

Please refer to our code

Lesson learnt

One-Class SVM vs Isolation Forest

We noticed while working on this project that the performance of One-Class SVM was not as good as Isolation Forest after we run the algorithms. After reading a few articles, we realized that in order for One-Class SVM to perform well, we needed a clean training dataset to train the model and set the boundary to detect outliers in the testing dataset.

Therefore, for our second iteration, we focused on training our model with data without any anomalies.

Questions?

References

1. https://www.comparitech.com/blog/information-security/canada-cyber-crime-statistics/
2. https://www.comparitech.com/blog/information-security/ddos-statistics-facts/
3. https://www.baeldung.com/cs/one-class-svm
4. https://www.analyticsvidhya.com/blog/2021/07/anomaly-detection-using-isolation-forest-a-complete-guide/
5. https://www.researchgate.net/figure/Isolation-Forest-learned-iForest-construction-for-toy-dataset_fig1_352017898
6. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/Z834IK
7. https://www.kaggle.com/datasets/julienjta/nyc-taxi-traffic