Hey Nick,

As far as you know, we have experienced an attack on our servers recently. To avoid further attacks, you were provided with a complete server's log to implement an anomaly detection using clustering algorithms and inform me of the results. I reviewed your working repository for this purpose last night. Your punctuality and robust implementation were appreciable, however, there are some drawbacks with your model that I would like to mention.

While implementing any machine learning algorithm, there are some hyperparameters you need to set in order to get the best performance. Examples of these hyperparameters in your implemented model are k (number of clusters) and t (threshold). Reviewing your code, I realized that you have hardcoded these values (k=8, t=0.97). Regarding these values, the question may arise: "Can we get better performance by changing these values?" Maybe.

To do so, you need to try different values of k and t to see if it can enhance your model performance. There are several approaches to determine these values. The first one is Grid Search where you manually define a subset of values for each hyperparameter, run your model using each pair and return the values for which the model performs best.

Another approach is Random Search in which instead of iterating through all items in a subset, you select some values randomly in a specific range and test your model using randomly selected values. You will select the values for which the model performed best.

The most successful approach, however, is Bayesian Optimization. In this approach, you should initially select a random point and evaluate model performance. Using this value, you need to build an underlying probabilistic function that shows expected improvements as well. From this function, you need to select the point that has maximum improvement. This is your next sample. Using this sample point evaluate the model performance again. Repeat the process iteratively until you make sure that you have found the desirable model performance. The advantage of this method over 2 before-mentioned methods is that it has no manual effort to define a range. Also, it leads us to the best performance in fewer iterations. I would like you to apply this optimization to your model and send me your final implementation.

Should you have any comment or question, please share with me via email or phone call.

Best Regards,
Shahram