Challenge Question 1

This submission is by Team BruteForce by Apoorv Walia, Gurjot Singh and Manu Maheshwari

The challenge provides us with 2 datasets.

The first dataset consists of measurements (real-values) of 41 sensors periodically (period of 1hr), generated for the duration of around one year. This data represents the behavior of 41 sensors collected periodically over a duration when there is no attack on the system. So they represent nominal behavior of the system. However, there is a chance that this data may be subject to data poisoning or false data injection attacks. Thus, while using this data to learn the nominal behavior of the system, you may have to use some robust machine learning technique.

The second dataset is collected from the same architecture but during seven different attacks. Each attack is active for a duration, as given below. This dataset aims to be useful to validate your model's performance. You may or may not use this dataset to do the training of your model.

Data PreProcessing

We drop the columns/sensor data which don't have different values in df(first dataset) and df2(second dataset). We split df into training and testing data in 8:2 ratio and standardise them using RobustScaler(). We reshape our data to resample it to 3d (sampole , n_past, n_future) and go on to build the main LSTM model using Keras.

LSTM Keras Model

We use 2 LSTM encoder layers and 2 LSTM decoder layers each with 50 Neurons. Add regularizers to mitigate overfitting. We use 60 epochs. We use 80% of our first datset to train the model.

After training we save out out model as 'keras_prob2.h5'

We use moving averages to compute smooth threshold for all features to mitigate the effect of outliers. We use the remianing 20% of our training data to compute Feature wise thresholds.

Visualise

We visualise model performance using various graphs which are detailed in the notebook FILE1.

We use DATASET2 to measure the performance of our model. We use plots to visualise it. Whereas, we use classification report and confusion matrix to look at the accuracy numerically.

Folder Guide

folder |

- IDS_test.py |
- FILE 1.ipynb |
- result.csv |
- keras_prob2.h5 |
- trans.pkl |
- Subdataset1.csv |
- Subdataset2.csv

FILE 1: We have the training model which also contains the commands to save our scaler and model. It also contains our tests for the model on subdataset2.

IDS_test.py : The main .py file that you can run from cmd.

```
python IDS_test.py absolute_path_for_test.csv_file
```

result.csv : Our results.

keras_prob2.h5 : Our model. Please don't get confused by the name.

trans.pkl: Our scaler

Subdataset1, Subdataset2 : csv files provided to us by committee

In [ ]: