AutoML in Cybersecurity - Activity 02

1 Purpose

This activity aims to leverage the knowledge we obtained from the previous activity and develop firsthand experience with AutoML techniques and tools. In particular, we are conducting an empirical study on the application of AutoML in Cybersecurity. The activity is divided into the following substacks:

a. Reading and summarizing selected articles on AutoML in Cybersecurity.

b. Identify benchmark datasets (or new datasets) for different applied machine learning cybersecurity challenges.

c. Select a set of AutoML tools to investigate their performance, with the datasets identified in subtask 2

2 Reading Papers

Things to do:

- Provide a critical (summary, strength and weakness, and possible improvement or extension) review for the selected papers on AutoML and Cybersecurity.
- Select 1 paper to discuss in details.
- Suggest additional articles to extend the literature
- Propose 2-3 research questions that needed to be answers.

3 Benchmark Datasets

Things to do:

- Identify 3-5 papers that apply ML in one of the following categories: malware detection, fraud detection, phishing and spam, data exfiltrations.
- Identify at least 2-3 datasets in each of the following categories: malware detection, fraud detection, phishing and spam, data exfiltrations.
- Provide a summary of the datasets and the performance of ML techniques used with theses datasets.

4 Empirical Study

Things to do:

- Based on subtasks 1 and 2, delevep research questions, problem statement.
- Develop a methodology to conduct an empirical study to address the problem statement.
- Purpose a timeline to complete the empirical study.

Useful Resources

- *How Kaggle solved a spam problem in 8 days using AutoML*, Will Cukierski, May 27, 2020 `https://cloud.google.com/blog/products/ai-machine-learning/how-kaggle-solved-a-spam-problem-using-automl`
- *Can you detect fraud from customer transactions?*, IEEE-CIS Fraud Detection `https://www.kaggle.com/c/ieee-fraud-detection/overview`
- *Automatic Machine Learning in Fraud Detection Using H2O AutoML*, Yuefeng Zhang, Nov 13, 2019 `https://towardsdatascience.com/automatic-machine-learning-in-fraud-detection-using-h2o-automl-6ba5cbf5c79b`
- *The TON_IoT data set*, by Dr Nour Moustafa,Oct 24, 2019 `https://research.unsw.edu.au/projects/toniot-data-set`
- *Aposemat IoT-23 A labeled dataset with malicious and benign IoT network traffic*, Sebastian Garcia, Agustin Parmisano, Maria Jose Erquiaga, 2020, `https://www.stratosphereips.org/datasets-iot23`