CREATIVE AND INNOVATIVE PROJECT

A User-Centric Machine Learning Framework for Cyber Security Operations Center

Project Documentation

G. Santhosh 2016103057
S. Ben Stewart 2016103513

Abstract

To assure cyber security of an enterprise, typically SIEM (Security Information and Event Management) system is in place to normalize security events from different preventive technologies and flag alerts. Analysts in the security operations center (SOC) investigate the alerts to decide if it is truly malicious or not. However, generally the number of alerts is overwhelming capacity to handle all alerts. Because of this, potential malicious attacks and compromised hosts may be missed. Machine learning is a viable approach to reduce the false positive rate and improve the productivity of SOC analysts. In this paper, we develop a user- centric machine learning framework for the cyber security operation center in real enterprise environment. We discuss the typical data sources in SOC, their workflow, and how to leverage and process these data sets to build an effective machine learning system. We use the system using the key repository of information regarding the vulnerabilities that allow intruders to breach computer networks is the National Vulnerability Database (NVD). NVD is a product of the U.S. National Institute of Standards and Technology's (NIST) Computer Security Division and is also sponsored by the U.S. Department of Homeland Security's Computer Emergency Readiness Team (US-CERT). We are implementing the below steps from data massaging, label creation, feature engineering, machine learning algorithm selection, model performance evaluations, to risk score generation.

The above implementation would help other teams with only knowledge of machine learning to get a better understanding of the domain of cyber security and the challenges it provides with the requirement of high accuracy models though the dataset is highly biased. It also helps the teams on the other side of the spectrum who are from the cyber security to get to understand how machine learning models can be used to the greater benfit.

Workflow

We will be using the NVD dataset as the base and create the labels for the records so that they could be used for the training process. Next, it would be necessary to get the most important features that would be required to find the anomalies with greater accuracy. Finally, we get to choose the best model required for the use case at hand. This model was tested against the Common Vulnerability Scoring System (CVSS) v3.0 standards as benchmark.