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

CREATIVE AND INNOVATIVE PROJECT REPORT

Submitted by

G. Santhosh (2016103057)

S. Ben Stewart (2016103513)

P. Udaykumar (2016103622)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

College of Engineering, Guindy

ANNA UNIVERSITY: CHENNAI 600 025

OCTOBER 2019

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project report "**A User-Centric Machine Learning Framework for Cyber Security Operations Center**" is the bonafide work of "G.Santhosh (2016103057), S.Ben Stewart (2016103513) and P.Udaykumar (2016103622)" who carried out the project work under my supervision.

Place: Chennai	**Dr. AR. Arunarani**
Date: 24/10/2019	**SUPERVISOR**

Teaching Fellow
Department of Computer Science and Engineering,
College of Engineering, Guindy
Anna University.

ACKNOWLEDGEMENTS

We would like to take this opportunity to thank the head of our department **Dr. S. Valli** who helped us with the lab and all issues we placed balancing the placement season with this semester. She went all the way to make sure the students were not stressed with the classes and projects.

We would also like to thank our supervisor **Dr. AR. Arunarani** for helping us to successfully complete this project and made sure to be there for us when we faced issues in moving the project forward.

We also have to thank **Ms. N. Kalaichelvi** for helping us to make our presentation and documentation free from bugs as the computer jargon would dictate.

This project would not have been possible if not for the tireless efforts by the members of this team in all the divisions in which they were asked to work on. Right from the project ideation till the implementation and the documentation and presentation part. We would also thank our mentors and friends who helped us when the project hit a few roadblocks.

Thanking you

G. Santhosh

S. Ben Stewart

P. Udaykumar

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 cyber security to get to understand how machine learning models can be used to the greater benefit.

TABLE OF CONTENTS

LIST OF FIGURES

REFERENCES

[1] Choudhury, S., & Bhowal, A. (2015). Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection. 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 89-95.

[2] Kaur, H. (2014). Algorithm used in Intrusion Detection Systems:a Review.

[3] SANS Technology Institute." The 6 Categories of Critical Log Information" 2013.

[4] X.Li and B.Liu."Learning to classify text using positive and unlabeled data", Proceedings of the 18th international joint conference on Artificial intelligence, 2003

[5] A user-centric machine learning framework for cyber security operations center https://ieeexplore.ieee.org/document/8004902

[6] Downloading and unzipping a .zip file without writing to disk https://stackoverflow.com/questions/5710867/downloading-and-unzipping-a-zip-file-without-writing-to-disk

[7] Parsing JSON Dataset with Pandas https://www.geeksforgeeks.org/pandas-parsing-json-dataset

[8] Filtering Pandas DataFrames on dates https://stackoverflow.com/questions/22898824/filtering-pandas-dataframes-on-dates

[9] Beautiful Soup to parse url to get another urls data https://stackoverflow.com/questions/4462061/beautiful-soup-to-parse-url-to-get-another-urls-data

[10] Adding a y-axis label to secondary y-axis in matplotlib https://stackoverflow.com/questions/14762181/adding-a-y-axis-label-to-secondary-y-axis-in-matplotlib

[11] Rotate axis text in python matplotlib https://stackoverflow.com/questions/10998621/rotate-axis-text-in-python-matplotlib

[12] Become a Machine Learning Engineer https://www.udacity.com/course/machine-learning-engineer-nanodegree--nd009t

[13] National Vulnerability DatabaseCVE-2017-5638 Detail https://nvd.nist.gov/vuln/detail/CVE-2017-5638