Real Time Threat hunting using Machine Learning Algorithms

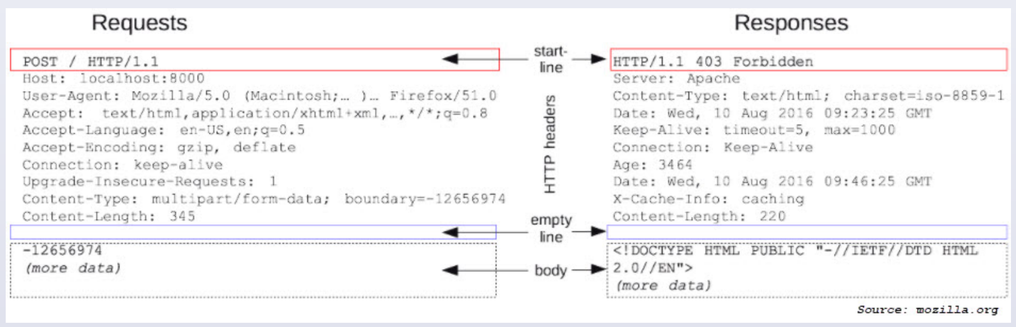
Student(s)
Elif Pınar Ön

Faculty Member(s)
Albert Levi

Sabancı Üniversitesi

PURE
PROGRAM FOR UNDERGRADUATE RESEARCH

Introduction

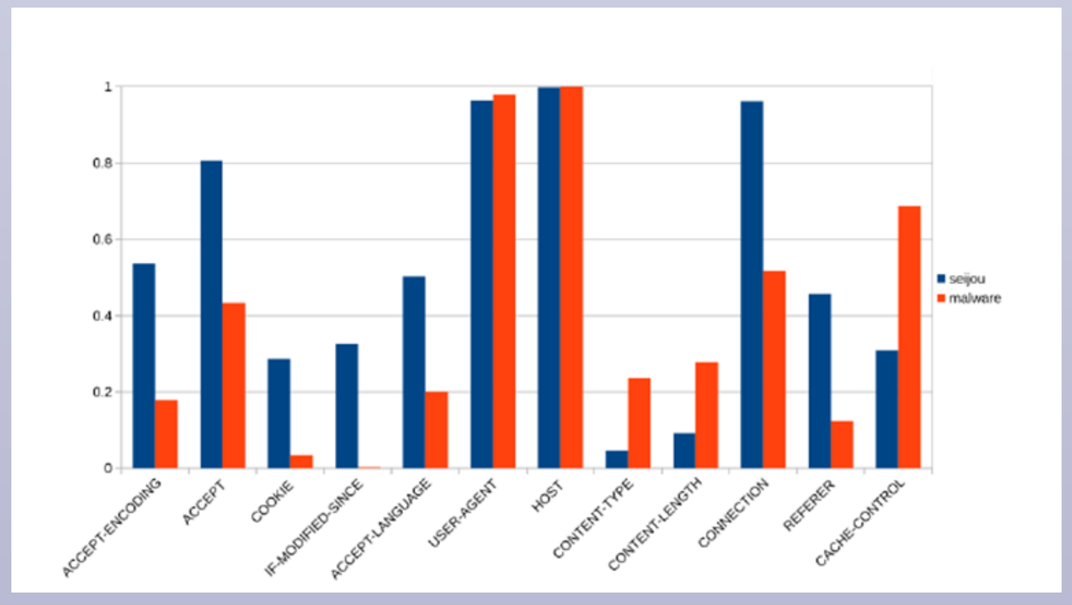


HTTP communication is done using HTTP headers which are the name or value pairs that are displayed in the request and response messages of message headers for Hypertext Transfer Protocol (HTTP). The main purpose of this project is to reduce the workforce of the Kuveyt Türk Katılım Bank workers by generating a malicious activities detecting mechanism which automates threat hunting. While automating the detection mechanism the main attribute of the machine learning models were HTTP header information. Some specific problems encountered while generating these models.

These problems can be listed as:

- Understanding the header information and classifying them.
- Finding labeled datasets for generating the detection models.
- If datasets couldn't be found, then separate tools would be used in order to create normal activities and malicious activities.
- Generated datasets will be labeled for classification methods of machine learning.
- Finding the best accuracy scored and appropriate machine learning algorithm for these datasets.

Research



Picture : Average of Client Header

Calderon, paul;hasegava,hirozaku;yamaguchi, yukiko;Shimada,hajime. Malware Detection based on HTTPS Characteristic via Machine Learning. Retrived from http://www.scitepress.org/Papers/2018/66546/66546.pdf

Header information differences between malicious activities and normal activities and the most significant attributes that affects the manner of the HTTP activity are;

"CONNECTION", "ACCEPT", "ACCEPT- ENCODING", "ACCEPT-LANGUAGE", "COOKIE", "CONTENT TYPE", "CACHE- CONTROLS", "IF- MODIFIED-SINCE"

Data Preparation & Tools

Ready to use HTTP header datasets were not proper for the problem, so unprocessed data in pcap format which are shared public on the Internet collected from the mentioned resources.

- Collection of pcap files includes malware collected from contagio blog whose owner is Mila Parkour.
- Collection of pcap files which are considered normal(without malware) collected from Stratosphere IPS dataset

The pcap files collected converted in to log files using Bro tool and modified Bro-Module. To manage collected dataset easily in the implementation part of machine learning algorithms, log files converted in to csv files and these csv files labeled as malicious or not, by using python.

What do we use?

- Normal-Crime PCAP files
- Bro ve Bro module
- Python, Anadonda
- Pandas ve sklearn
- Multinomial Naive Bayes

ANACONDA®

BRO NETWORK SECURITY MONITOR

python

Implemention of Machine Learning Algorithm

Multinomial Naive Bayes algorithm has been chosen according to past researches. However, these algorithms are working with integers and our dataset content is string. Thus, Count Vectorizers used in order to convert these strings to integers. Count Vectorizer is converting these strings in to vectors that include absence of strings. In order to use Count Vectorizer, Bro-Module has been modified and the mentioned significant attributes combined in to a single column.

Conclusion

Confusion Matrix	
True Negative	2759
False Positive	26
False Negative	6
True Positive	4257

We have 35238 different HTTP activities in our dataset, and we used 80% of this data for training models and 20% of this data for testing the models and we have the results mentioned above.

The results were promising, only Multinomial Naive Bayes algorithm has been implemented and as expected count vectorizer with the attributes considered as significant to distinguish malware and normal headers resulted in a very low False Negative outcomes. The most dangerous mistake of the model would be the False Negative predictions because the main aspect of using a machine learning model is to lower the workforce of threat hunting, if the algorithm would have a high mistake rate on false negatives the results would be misleading.

"This project is a SOP project of Kuveyt Türk Katılım Bank."

KUVEYTTÜRK
KATILIM BANKASI A.Ş.

References

- Threat Hunting with Python: Prologue and Basic HTTP Hunting. (2017, September 18). Retrieved from, https:// dgunter.com/2017/09/17/threat-hunting-with-python- prologue-and-basic-http-hunting/
- Threat Hunting for HTTP User Agents | Sqrrl. (n.d.). Retrieved from https://sqrrl.com/ threat-hunting-http-user-agents/
- Threat Hunting vs Incident Response: Getting Proactive Instead of Staying Reactive. (2018, October 10). Retrieved from https://dgunter.com/2018/10/09/threat-hunting-vs- incident-response-getting- proactive-instead-of-staying-reactive/
- Zegers/University of Amsterdam, R. (2015). *HTTP Header Analysis*.
- Stevanovic/Aalborg University, M., & Pedersen/Aalborg University, J. (2015). *On the Use of Machine Learning for Identifying Botnet Network Traffic*.
- http://contagiodump.blogspot.com/2013/04/collection-of-pcap-files-from-malware.html
- https://www.stratosphereips.org/datasets-normal/