Malicious URLs Detection: A Machine Learning Approach

Yash Choksi, Tanvi Gawade, Aakanksha Tasgaonkar
Faculty Supervisor: Dr. Shilpa Balan
California State University, Los Angeles

Introduction

- With the worldwide use of Internet technology, the concern with security becomes the priority.
- Web applications all around the world have become popular and bring people convenience, while there is a rapid growth in the number of attacks from various criminal enterprises, such as financial fraud and spam-advertised commerce [1].
- This research study aims to analyze classification approach for different URL attack types, namely, spam and malware.

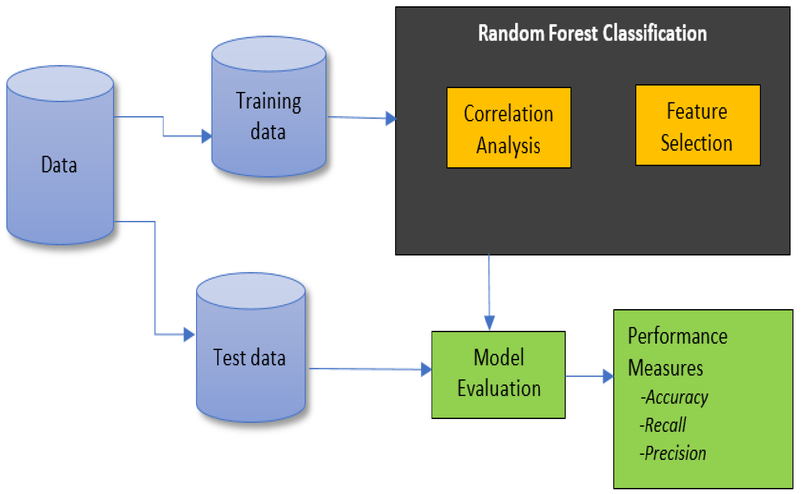
Background

- Visitors of web URLs are under the threat of being a victim of certain attacks [2].
- Malicious website is a common and serious threat to cybersecurity [3].
- Malicious URLs host unsolicited content and trap unsuspecting users to become victims of scams (monetary loss, theft of private information, and malware installation), and cause losses of billions of dollars every year [3].
- 24,000 malicious applications are blocked every day, and information that most applications release is 63% mobile phone numbers and 37% device location [4].
- From 2016 to 2017 the percentage of cybersecurity costs increased by 22.7%, with malware attack costing companies 2.4 million dollars on average [4].
- Recent statistics show that on an average, 8% tweets consist of spam and other malicious content [5].

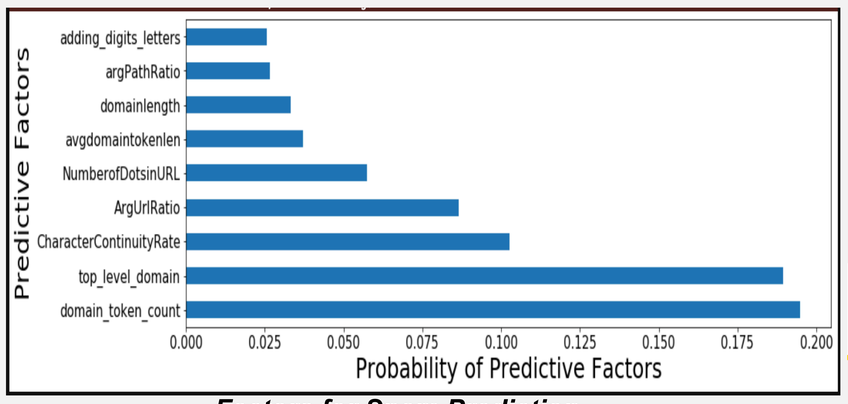
Methodology

- In this study, we analyzed the data set available, i.e. WEBSPAM-UK2007 dataset [6] which consists of 12,000 spam URLs and 11,500 URLs pertaining to malware websites.
- The Malware URLs dataset were gathered from DNS-BH which is a project that maintains list of malware sites.
- We analyzed the datasets containing spam and malware URLs and wrote a Python script to perform the feature selection and created a predictive model using Random Forest Classification [7].
- Random Forest Classification, a machine learning technique is applied in this study to classify malicious websites.

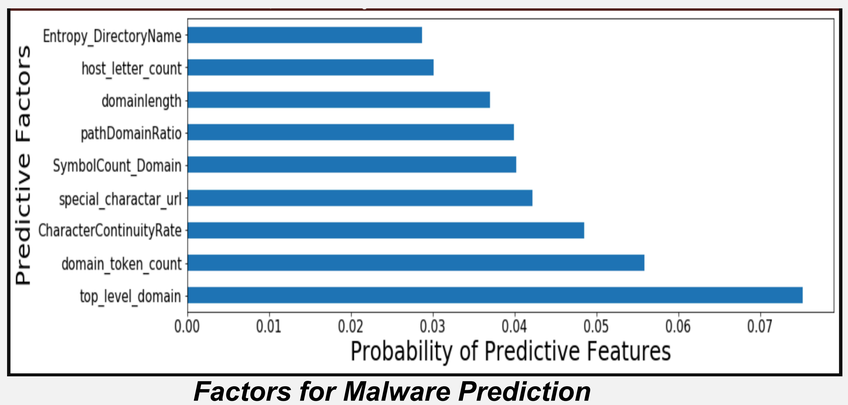
Research Model



Analysis

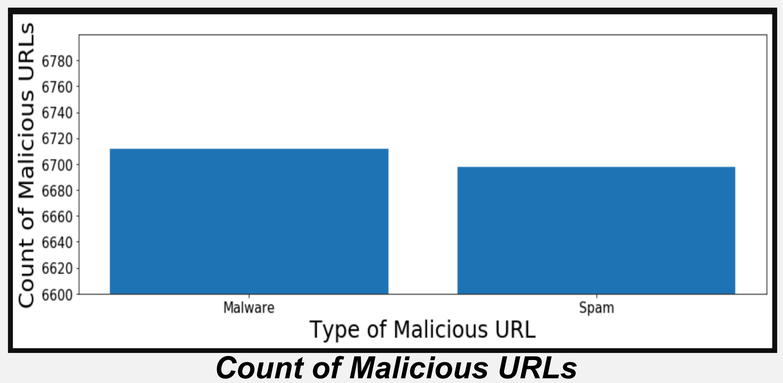


Factors for Spam Prediction



Factors for Malware Prediction

Analysis



Count of Malicious URLs

Recall and Precision

$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Category	Recall	Precision	Accuracy
Spam	0.998	0.99	0.99
Malware	0.99	0.996	0.99

References

[1] Liu, H., Pan, X., Qu, Z. (2019). Learning Based Malicious Web Sites using Suspicious URLs. ACM.
[2] Mamun, Mohammad & Rathore, Muhammad & Habibi Lashkari, Arash & Stakhanova, Natalia & Ghorbani, Ali. (2016). Detecting Malicious URLs Using Lexical Analysis. 9955. 467-482.
[3] Sahoo, D., Liu, C., Hoi, S. (2019). Malicious URL Detection using Machine Learning: A Survey. ACM.
[4] Ferreira, M. (2019). Malicious URL Detection using Machine Learning Algorithms. Proceedings of the Digital Privacy and Security Conference. pp. 114-122.
[5] Chaudhari, R., Dakhane, D. (2016). Machine Learning Approach for Detection of Malicious URLs and Spam in Social Network. IRJET. pp. 835-839.
[6] Web Spam data set (2007). Available at http://chato.cl/webspam/datasets/uk2007/
[7] Yiu, T. (2019). Understanding Random Forest. How the Algorithm Works and Why it is so Effective. Toward Data Science. Available at https://towardsdatascience.com/understanding-random-forest-58381e0602d2