BDA Project: Malicious And Benign Website URL Detection

Nguyen Xuan Binh

1st February 2023

Contents

Introduction

1. Central problem

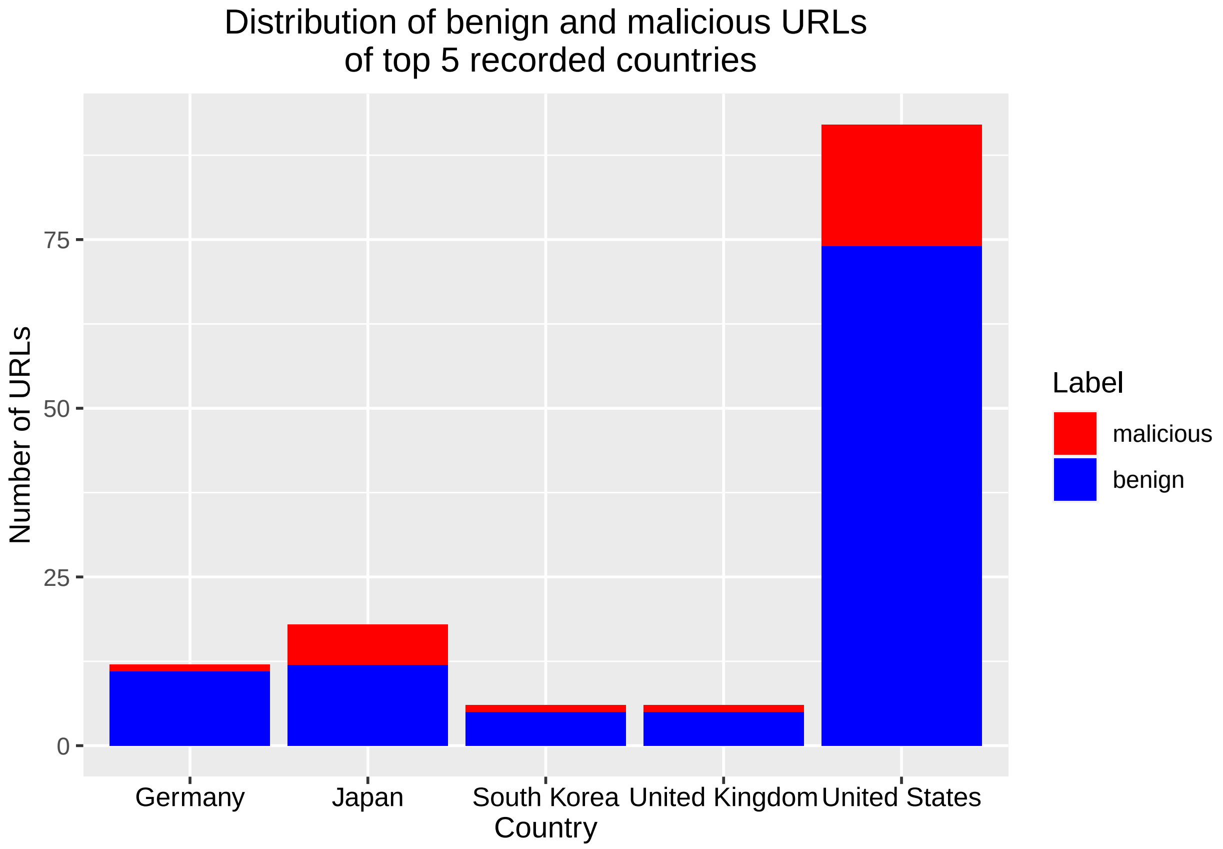
Detection of malicious URLs among the benign ones is a crucial goal of modern-day cybersecurity as it helps prevent individuals and organizations from falling victim to phishing, data breaching, malware infections, and other types of cyber threats. The most common type is phishing, where the URLs are disguised as valid sites to trick users into revealing their credentials. Some other types even install harmful softwares or redirect users to other malicious sites. With the rapid growth of the internet and the increasing dependence on technology, black-hat hackers and thieves have found innovative ways to spread their malicious content through fake URLs. A 2017 report from Cybersecurity Ventures predicted ransomware damages would cost the world \$5 billion in 2017, up from \$325 million in 2015 — a 15X increase in just two years. The damages for 2018 were predicted to reach \$8 billion, and for 2019 the figure is \$11.5 billion (Morgan (2019)). Therefore, it is an urgent task to automate the process of detecting and blocking malicious URLs floating on the net.

2. Motivation

In order to protect against these threats, it is essential to detect malicious URLs and prevent individuals and organizations from accessing them. This can be accomplished through various techniques, including URL reputation analysis, machine learning algorithms, and network security solutions. By detecting and blocking malicious URLs, individuals and organizations can better protect themselves and their sensitive information from cyber-attacks. In this report, I aim to detect malicious URLs among the benign or safe ones based on various features of the URLs and the websites associated with them. The analysis method will be based on the Bayesian inference approach to account for past data on the recorded URLs.

3. Main modeling idea

The problem is the detection of malicious URLs, which means it is a classification task. As a result, I will heavily use the Beta distribution to model the probabilities for each feature and the Bernoulli distribution for modeling the likelihood of both labels and features. Based on intuition, the rate of malicious URLs is expected to vary depending on the countries and regions. For example, reputable countries in cybersecurity, such as Finland and UK, will host much fewer malicious domains/URLs. In contrast, others, such as Russia, China, and Vietnam, are less regulated and will have more harmful network content. From this belief, I decided to split the recorded URLs depending on the countries and proceeded to perform two Bayesian models: the separate model, where the rates of malicious URLs from each country are individually analyzed, and the pooled model, where all URLs are merged and treated as if they come from only one source. Both models are equally valid in that the separated model looks from the perspective of regional difference, while the pooled model looks from the common origin on the Domain Name System (DNS). Below is the illustration of malicious and benign URLs distribution among the recorded countries.



Dataset

1. Data description

The dataset in this report is collected by an author named A.K.Singh in his research paper for International Conference on Communication Systems & Networks (A. K. Singh and Goyal (2019)). This dataset specifically caters for machine learning-based classification analysis of malicious and benign webpages. According to the author, this dataset comprises of various extracted attributes and raw webpage content, which are:

- 'url' - (string) The URL of the webpage
- 'ip_add' - (string) IP address of the webpage.
- 'geo_loc' - (string - categorical) The geographic location where the webpage is hosted.
- 'url_len' - (float) The length of URL.
- 'js_len' - (float) Length of JavaScript code on the webpage.
- 'js_obf_len - (float) Length of obfuscated JavaScript code.
- 'tld' - (string - categorical) The top level domain of the webpage.
- 'who_is' - (binary) Whether the WHO IS domain information is complete or not.
- 'https' - (binary) Whether the site uses https or http.
- 'content' - (string) The raw webpage content including JavaScript code.
- 'label' - (binary) The class label for benign or malicious webpage.

Because his dataset is extremely heavy and extensive (1.3 million datapoints for training data and nearly 340000 datapoints for testing data), I only extract a tiny portion from it, which is 120 training and 250 testing datapoints to allow the Stan sampling to run adequately fast.

2. Data source and analysis difference

The source of the dataset can be found at:

Data source description (A. K. Singh (2020)): https://data.mendeley.com/datasets/gdx3pkwp47/2

Data source download website: https://www.researchgate.net/publication/347936136_Malicious_and_Benign_Webpages_Dataset

It is also available on Kaggle: https://www.kaggle.com/datasets/aksingh2411/dataset-of-malicious-and-benign-webpages

The difference between this report and the paper of A.K.Singh is that he focuses on comparing various machine learning strategies to tackle this problem, which are supervised and unsupervised learning, while this report solely focuses on the Bayesian inference technique, combined with supervised learning to predict the malicious URLs. His paper did mention a Bayesian technique, but it is Naive Bayes Classifier, while the Bayesian technique in this project is based on probabilistic sampling in separate and pooled models.

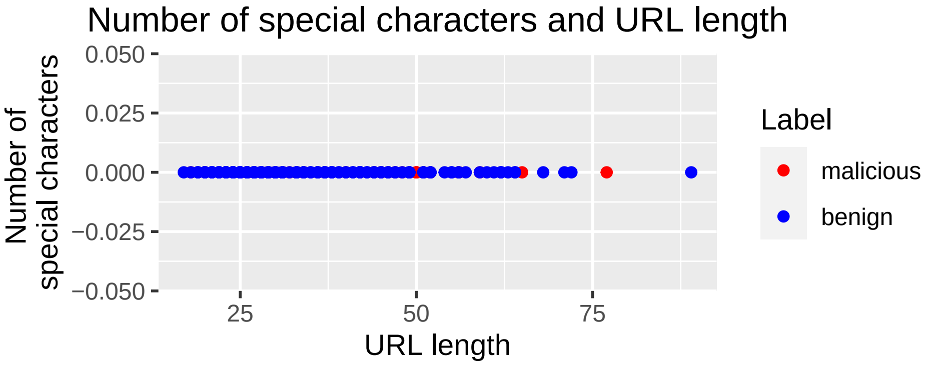
3. Feature selection and data cleaning

Feature selection is crucial to dimension reduction and model accuracy and runtime improvement. I have set two criteria: the number of features should be at most four, and the features should be highly correlated with the label (malicious/benign).

First is the URL itself. Based on the URL alone, it is hard to determine whether it is related to the underneath danger, so I decided to extract the number of special characters from the URL. This is the original dataset after I calculated the num_special column

```
head(test_websites)
```

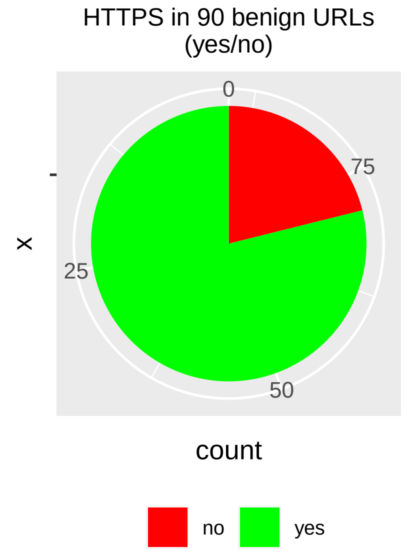
```
##   X label url_len       geo_loc https js_len js_obf_len    who_is num_special
## 1 1   bad      34 United States    no  324.0      0.000 incomplete           0
## 2 2   bad      50         China    no  449.1    273.951 incomplete           0
## 3 3   bad      29         India    no  484.2    329.256 incomplete           0
## 4 4   bad      46        Turkey    no  745.2    514.188 incomplete           0
## 5 5   bad      56       Germany    no  432.0    194.400 incomplete           0
## 6 6   bad      31     Argentina    no  393.3    275.310 incomplete           0
```

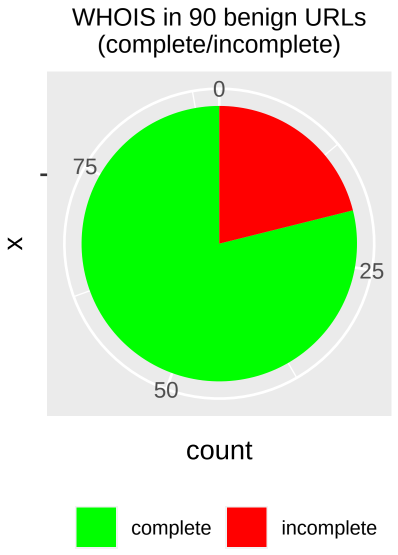


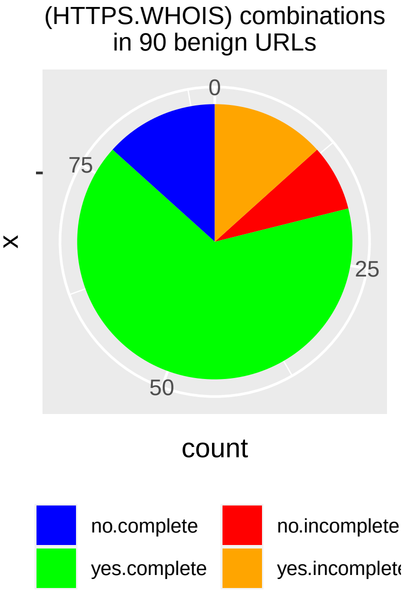
It appears that the URL name itself is not helpful for prediction, including its length. This can seen from the graph above as all malicious and benign URL lengths are randomly distributed, while the number of special characters are all 0s. Therefore, I omitted the features “url” and “url_len.”

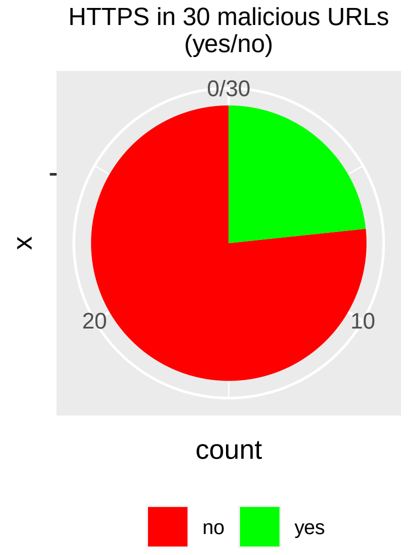
Next is the top level domain (tld) name. While some domains may be notoriously dangerous, such as .zip, .link and .review, most of the tld in the dataset are the most popular domains, such as .com, .org and .net. All of them are equally likely to be benign or malicious as they are prevalent on WWW. As a result, I omitted tld feature as it is not particularly helpful in prediction.

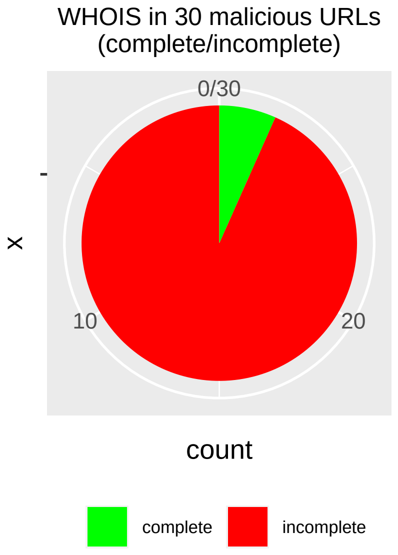
Next, I examine two features: https and whois. HTTPS (Hypertext Transfer Protocol Secure) is an extension of the HTTP. It uses encryption for secure communication over a computer network, and is widely used on the Internet. As a result, websites with https are much more likely to be secure and safe compared to http websites. On the other hand, WHOIS is a query and response protocol that is widely used for querying databases that store the registered users or assignees of an Internet resource, such as a domain name, an IP address block or an autonomous system. Websites with completed whois registration is much safer and transparent than unregistered websites. Therefore, I decided to keep these two features.

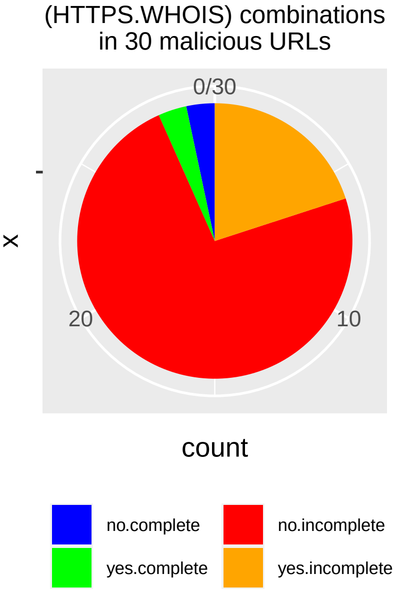

HTTPS in 90 benign URLs (yes/no)


WHOIS in 90 benign URLs (complete/incomplete)


(HTTPS.WHOIS) combinations in 90 benign URLs

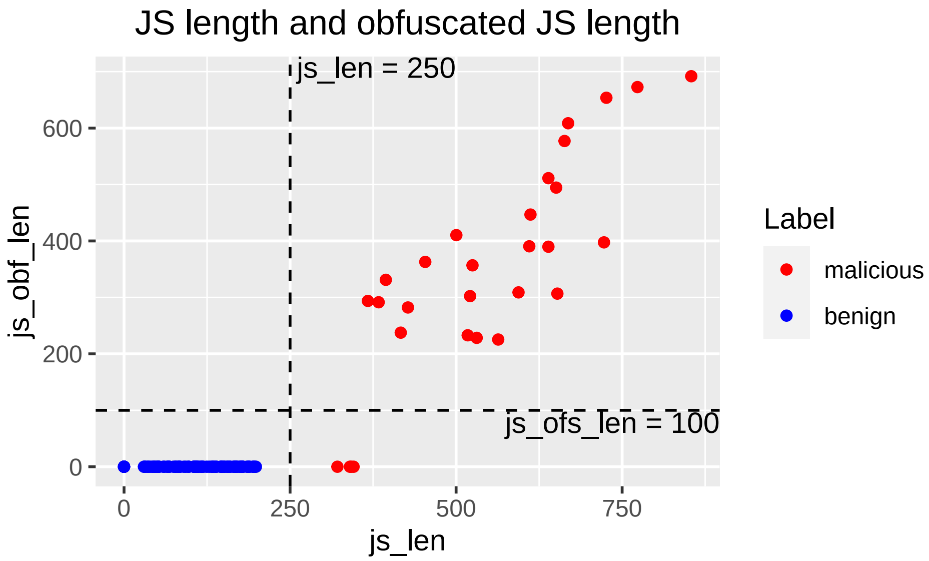

HTTPS in 30 malicious URLs (yes/no)


WHOIS in 30 malicious URLs (complete/incomplete)


(HTTPS.WHOIS) combinations in 30 malicious URLs

From the pie charts, it is evident that https and completed whois registration are strongly related to the label. Most benign websites have https and completed whois but otherwise for malicious ones.

Finally, two features left are the Javascript length and the obfuscated Javascript length of the web raw content. Javascript is the native language of web browsers and inherent to all running websites. Because Javascript code is open-source by Inspection, some organizations want to prevent others to copy their code. Therefore, JavaScript obfuscation is a series of code transformations that turn plain, easy-to-read JS code into a modified version that is extremely hard to understand and reverse-engineer.



When plotting them together, a clear pattern has arisen. It appears that all URLs having JS length longer than 250 are malicious and benign when smaller than 250. Regarding the obfuscated JS length, if it is larger than 100, the URL is almost certain to be malicious. If it is smaller than 100, the URL is likely to be benign, as the number of red points are much smaller than the blue points under the line js_obf_len = 100. Observing this distinction, I decided to transform these two float features into binary formats as follows:
- $js_{len} \geq 250 => js_{len\_binary} = 1$ and 0 otherwise
- $js_{obf\_len} \geq 100 => js_{len\_obf\_binary} = 1$ and 0 otherwise

Because the https, whois and label columns are in string formats, I also need to convert them to binary formats (0/1) so that it can be passed into the Stan models. The binary labels are:
- $label = ”good” => label\_bin = 0$ and $label = ”bad” => label\_bin = 1$
- $https = ”yes” => https\_bin = 0$ and $https = ”no” => https\_bin = 1$
- $whois = ”complete” => whois\_bin = 0$ and $whois = ”incomplete” => whois\_bin = 1$

From this binary format, it appears that js_len and js_obf_len have inverse proportion while https and whois have direct proportional to the label according to the analysis above. In total, there are four features in this report: https, whois, js_len and js_obf_len. Finally, the geo_loc column indicates which country the URL originates from. It is used to partition the URLs into different countries for the separate and pooled models. As a result, geo_loc is not a feature in this report. The cleaned dataframe now becomes:.

```
##   label_bin       geo_loc https_bin whois_bin js_len_bin js_obf_len_bin
## 1         1         China         0         1          1              1
## 2         1 United States         1         1          1              1
## 3         1       Germany         1         1          1              1
## 4         1 United States         1         1          1              1
## 5         1 United States         1         1          0              0
## 6         1         China         1         1          1              0
```

Separate model

1. Model description

2. Prior choice and justifications

3. Stan code and running options

4. Convergence diagnostics

5. Posterior predictive checks

6. Predictive performance assessment

7. Prior sensitivity analysis

Pooled model

1. Model description

2. Prior choice and justifications

3. Stan code and running options

4. Convergence diagnostics

5.Posterior predictive checks

6. Predictive performance assessment

7. Prior sensitivity analysis

Model comparison

Discussion

- Existing issues
-
-
- Potential improvements
-
-

Conclusion

Reflection

References

Morgan, Steve. 2019. 2019. https://cybersecurityventures.com/global-ransomware-damage-costs-predicted-to-reach-20-billion-usd-by-2021/.

Singh, A K, and Navneet Goyal. 2019. "A Comparison of Machine Learning Attributes for Detecting Malicious Websites." In *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, 352–58. https://doi.org/10.1109/COMSNETS.2019.8711133.

Singh, A. K. 2020. "Malicious and Benign Webpages Dataset." *Data in Brief* 32: 106304. https://doi.org/https://doi.org/10.1016/j.dib.2020.106304.