Security Laboratory

Mini Project Report

July to October 2018

Phishing Detection

Instructor: Dr. Arockia Xavier Annie R, Asst. Prof. DCSE, CEG.

Team:
Sangameswaran R S – 2015103607
Krishna Gapileswar V R – 2015103513

Abstract:

The primary objective of this mini project is to achieve Phishing detection. Phishing is a cybercrime in which a target or targets are contacted by email, telephone or text message by someone posing as a legitimate institution to lure individuals into providing sensitive data such as personally identifiable information, banking and credit card details, and passwords. The information is used across important accounts to thieve identity and cause financial loss. In most of the cases, the target is lured to follow a link which appears to be legitimate. Our project solves the risk of following that link by classifying the link as phished or not. The Random Forest classification algorithm is used. Fifteen features have been used for the classification and the accuracy of the classification is 94%.

Data Set:

One of the major challenges in implementing machine learning in cyber security is the availability of reliable dataset. The UCI Phishing Websites Data Set is used for this project.

Feature Set:

The following list of features have been used,

- IP address in URL
- URL length
- Tiny URL
- Presence of @ symbol
- Last index of //
- hyphen in URL domain
- Subdomain count
- HTTPS usage
- Domain registration length
- Favicon domain check
- Using non standard port
- HTTPS is URL domain part
- Request URL
- Request URL in anchor tag
- Having iframe

About the features:

IP address in URL:

If an IP address is used as an alternative of the domain name in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal information. Sometimes, the IP address is even transformed into hexadecimal

code as shown in the following link "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

Rule: IF $\begin{cases} If\ The\ Domain\ Part\ has\ an\ IP\ Address \rightarrow Phishing \\ Otherwise \rightarrow Legitimate \end{cases}$

URL Length:

Phishers can use long URL to hide the doubtful part in the address bar. For example:

http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html

To ensure accuracy of our study, we calculated the length of URLs in the dataset and produced an average URL length. The results showed that if the length of the URL is greater than or equal 54 characters then the URL classified as phishing. By reviewing our dataset we were able to find 1220 URLs lengths equals to 54 or more which constitute 48.8% of the total dataset size.

Rule: IF $\begin{cases} URL\ length < 54 \rightarrow feature = Legitimate \\ else\ if\ URL\ length \geq 54 \wedge \leq 75 \rightarrow feature = Suspicious \\ otherwise \rightarrow feature = Phishing \end{cases}$

Tiny URL:

URL shortening is a method on the "World Wide Web" in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an "HTTP Redirect" on a domain name that is short, which links to the webpage that has a long URL. For example, the URL "http://portal.hud.ac.uk/" can be shortened to "bit.ly/19DXSk4".

Rule: IF $\begin{cases} TinyURL \rightarrow Phishing \\ Otherwise \rightarrow Legitimate \end{cases}$

Presence of @ symbol*:*

Using "@" symbol in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol.

Rule: IF $\begin{cases} Url\ Having\ @\ Symbol \rightarrow Phishing \\ Otherwise \rightarrow Legitimate \end{cases}$

Last index of //:

The existence of "//" within the URL path means that the user will be redirected to another website. An example of such URL's is: "http://www.legitimate.com//http://www.phishing.com". We examin the location where the "//" appears. We find that if the URL starts with "HTTP", that means the "//" should

appear in the sixth position. However, if the URL employs “HTTPS” then the “//” should appear in seventh position.

$$\textit{Rule: IF} \begin{cases} \textit{ThePosition of the Last Occurrence of } /¿ \in \textit{the URL} > 7 \rightarrow \textit{Phishing} \\ \textit{Otherwise} \rightarrow \textit{Legitimate} \end{cases}$$

hyphen in URL domain:

The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example http://www.Confirme-paypal.com/.

$$\textit{Rule: IF} \begin{array}{c} -¿\textit{Symbol} \rightarrow \textit{Phishing} \\ ¿ \\ ¿ \\ \textit{Domain Name Part Includes} ¿ \\ ¿ \end{array}$$

Subdomain count:

Let us assume we have the following link: http://www.hud.ac.uk/students/. A domain name might include the country-code top-level domains (ccTLD), which in our example is “uk”. The “ac” part is shorthand for “academic”, the combined “ac.uk” is called a second-level domain (SLD) and “hud” is the actual name of the domain. To produce a rule for extracting this feature, we firstly have to omit the (www.) from the URL which is in fact a sub domain in itself. Then, we have to remove the (ccTLD) if it exists. Finally, we count the remaining dots. If the number of dots is greater than one, then the URL is classified as “Suspicious” since it has one sub domain. However, if the dots are greater than two, it is classified as “Phishing” since it will have multiple sub domains. Otherwise, if the URL has no sub domains, we will assign “Legitimate” to the feature.

$$\textit{Rule: IF} \begin{cases} \textit{Dots} \in \textit{Domain Part} = 1 \rightarrow \textit{Legitimate} \\ \textit{Dots} \in \textit{Domain Part} = 2 \rightarrow \textit{Suspicious} \\ \textit{Otherwise} \rightarrow \textit{Phishing} \end{cases}$$

HTTPS Usage:

The existence of HTTPS is very important in giving the impression of website legitimacy, but this is clearly not enough. The authors in (Mohammad, Thabtah and McCluskey 2012) (Mohammad, Thabtah and McCluskey 2013) suggest checking the certificate assigned with HTTPS including the extent of the trust certificate issuer, and the certificate age. Certificate Authorities that are consistently listed among the top trustworthy names include: “GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster and VeriSign”. Furthermore, by testing out our datasets, we find that the minimum age of a reputable certificate is two years.

$$\textit{Rule: IF} \begin{cases} \textit{Use https} \wedge \textit{Issuer Is Trusted} \wedge ¿\textit{ Age of Certificate} \geq 1 \textit{ Years} \rightarrow \textit{Legitimate} \\ \textit{Using https} \wedge \textit{Issuer Is Not Trusted} \rightarrow \textit{Suspicious} \\ \textit{Otherwise} \rightarrow \textit{Phishing} \end{cases}$$

Domain length:

Based on the fact that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.

$$\textit{Rule: IF} \begin{cases} \textit{Domains Expires on} \leq 1 \textit{ years} \rightarrow \textit{Phishing} \\ \textit{Otherwise} \rightarrow \textit{Legitimate} \end{cases}$$

Favicon domain check:

A favicon is a graphic image (icon) associated with a specific webpage. Many existing user agents such as graphical browsers and newsreaders show favicon as a visual reminder of the website identity in the address bar. If the favicon is loaded from a domain other than that shown in the address bar, then the webpage is likely to be considered a Phishing attempt.

Rule: IF $\begin{cases} Favicon\ Loaded\ Domain \rightarrow Phishing \\ Otherwise \rightarrow Legitimate \end{cases}$

Using non standard port:

This feature is useful in validating if a particular service (e.g. HTTP) is up or down on a specific server. In the aim of controlling intrusions, it is much better to merely open ports that you need. Several firewalls, Proxy and Network Address Translation (NAT) servers will, by default, block all or most of the ports and only open the ones selected. If all ports are open, phishers can run almost any service they want and as a result, user information is threatened. The most important ports and their preferred status are shown in Table 2.

Rule: IF $\begin{cases} \text{Port \# is of the } Preffered\ Status \rightarrow Phishing \\ Otherwise \rightarrow Legitimate \end{cases}$

PORT	Service	Meaning	Preferred Status
21	FTP	Transfer files from one host to another	Close
22	SSH	Secure File Transfer Protocol	Close
23	Telnet	provide a bidirectional interactive text-oriented communication	Close
80	HTTP	Hyper test transfer protocol	Open
443	HTTPS	Hypertext transfer protocol secured	Open
445	SMB	Providing shared access to files, printers, serial ports	Close
1433	MSSQL	Store and retrieve data as requested by other software applications	Close
1521	ORACLE	Access oracle database from web.	Close
3306	MySQL	Access MySQL database from web.	Close
3389	Remote Desktop	allow remote access and remote collaboration	Close

Common port numbers to be checked

HTTPS is URL domain part:

The phishers may add the “HTTPS” token to the domain part of a URL in order to trick users. For example, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.

Rule: IF $\begin{cases} \text{Using } HTTP\ Token \in Domain\ Part\ of\ The\ URL \rightarrow Phishing \\ Otherwise \rightarrow Legitimate \end{cases}$

Request URL:

Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate

webpages, the webpage address and most of objects embedded within the webpage are sharing the same domain.

$$\textit{Rule: IF} \begin{cases} of\ Request\ URL < 22 \rightarrow Legitimate \\ \%of\ Request\ URL \geq 22 \wedge 61 \rightarrow Suspicious \\ \quad Otherwise \rightarrow feature = Phishing \end{cases}$$

Request URL in anchor tag:

An anchor is an element defined by the <a> tag. This feature is treated exactly as "Request URL". However, for this feature we examine:

If the <a> tags and the website have different domain names. This is similar to request URL feature.

If the anchor does not link to any webpage, e.g.:

A) <a href="#">

B) <a href="#content">

C) <a href="#skip">

D) <a href="JavaScript ::void(0)">

$$\textit{Rule}: \text{ IF} \begin{cases} of\ URL\ Of\ Anchor < 31 \rightarrow Legitimate \\ of\ URL\ Of\ Anchor \geq 31 \wedge \leq 67 \rightarrow Suspicious \\ \quad Otherwise \rightarrow Phishing \end{cases}$$

Having iframe:

IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the "iframe" tag and make it invisible i.e. without frame borders. In this regard, phishers make use of the "frameBorder" attribute which causes the browser to render a visual delineation.

$$\textit{Rule: IF} \begin{cases} Using\ iframe \rightarrow Phishing \\ Otherwise \rightarrow Legitimate \end{cases}$$

Random Forests:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.Random decision forests correct for decision trees' habit of overfitting to their training set.

Implementation Language: Python

Github repo link: https://github.com/SangameswaranRS/PhishingDetection

ReferenceRepository:https://archive.ics.uci.edu/ml/machine-learning-databases/00327/