Literature Survey

Team ID : **IBM-Project-41888-1660645823**

Project Name: Web phishing Detection

College Name : The Kavery engineering college

Department : Computer Science And Engineering

Team Leader : Hari prathap R

Team Member : Arivuselvam P

Team Member : Nandhakumar

Team Member : Sanjay

Phishing Websites Features

Rami M.Mohammad
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK.

Fadi Thabtah
E-Business Department
Canadian University of Dubai

Lee McCluskey
School of Computing and Engineering
University of Huddersfield
Huddersfield, UK.

1.Phishing Websites Features:

One of the challenges faced by our research was the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using data mining techniques have been disseminated these days, no reliable training dataset has been published publically, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features.

In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we proposed some new features, experimentally assigned new rules to some well-known features and updated some other features.

2.Address Bar based Features:

Using the IP Address

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{ Url ip Address>>>phishing
otherwise>>>legitimate

Long URL to Hide the Suspicious Part:

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.

Redirecting using "//"

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 examine 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.

Rule: IF

Adding Prefix or Suffix Separated by (-) to the 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

Sub Domain and Multi Sub Domains

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 "had" 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.

HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)

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.

Domain Registration 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.

Table 1 Common ports to be checked

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 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

Abstract

The detection of phishing attacks. Phishing attacks target vulnerabilities that exist in systems due to the human factor. Many cyber attacks are spread via mechanisms that exploit weaknesses found in end-users, which makes users the weakest element in the security chain. The phishing problem is broad and no single silver-bullet solution exists to mitigate all the vulnerabilities effectively, thus multiple techniques are often implemented to mitigate specific attacks.

This paper surveys the features used for detection and detection techniques using machine learning. Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack.

The Main aim of the attacker is to steal banks account credentials. Phishing attacks are becoming successful because of lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

Phishing maybe a style of broad extortion that happens once a pernicious web site acts sort of a real one memory that the last word objective to accumulate unstable info, as an example, passwords, account focal points, or MasterCard numbers. all the same, the means that there square measure some of contrary to phishing programming

ALGORITHMS USED :

Two algorithms have been implemented to check whether a URL is legitimate or fraudulent.

Random forest algorithm creates the forest with number of decision trees. A High number of trees gives high detection accuracy. Creation of trees is based on the bootstrap method. In the bootstrap method, features and samples of dataset are randomly selected with replacement to construct a single tree.

Decision tree begins its work by choosing the best splitters from the available attributes for classification which is considered as a root of the tree. The algorithm continues to build tree until it finds the leaf node. Decision tree creates training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to class label.

The accuracy of the model :

Research demonstrates that current phishing detection technologies have an accuracy rate **between 70% and 92.52%**. The experimental results prove that the accuracy rate of our proposed model can yield up to 95%, which is higher than the current technologies for phishing website detection.

Phishing Website Detection Using Machine Learning Algorithms:

PROPOSED WORK:

URLs extracting and analyze Various links by checking with Back listing with the help of Machine Learning to increase accuracy.

TOOLS USED/ ALGORITHM:

- Decision Tree Algorithm
- Random Forest Algorithm
- Support Vector Machine Algorithm

TECHNOLOGY:

- Machine Learning

ADVANTAGES/ DISADVANTAGES:

The disadvantage is that the Characteristics are not guaranteed to always exist in

such attacks and the false positive rate in detection is very high.
Advantage is 97.14% detection accuracy using random forest algorithm with lowest false positive rate**.**

Detecting phishing websites using machine learning technique:

PROPOSED WORK:

URL-based anti-phishing machine learning and method URL Net, a CNN-based deep-neural URL detection network.

TOOLS USED/ALGORITHM:

- Support Vector Machine
- K-NN
- Random forest classification
- Artificial Neural Network

TECHNOLOGY:

- Machine Learning

ADVANTAGES/ DISADVANTAGES:

Advantages-Reduces over fitting in decision trees and helps to improve the accuracy.

Disadvantages-Requires a computational power as well as resources as it builds numerous trees to combine their outputs.

Phishing Website detection using machine learning and deep learning techniques:

PROPOSED WORK:

It discusses the machine learning and deep learning algorithms and apply all these algorithms on our dataset and the best algorithm having the best precision and accuracy is selected for the phishing website detection.

TOOLS USED/ALGORITHM:

- •Regression Techniques
- •K nearest neighbor
- •Decision Tree
- •Random Forest
- •XG Boost
- •AdaBoost.

TECHNOLOGY:

- •Machine Learning , Deep Learning

ADVANTAGES/ DISADVANTAGES:

Advantage is to eliminate the cyber threat risk level. Increase user alertness to phishing risks.

Disadvantage is Negative effects on a business, including of money, loss of intellectual property

Phishing Website Detection Based on URL:

PROPOSED WORK:

To preserve the confidentiality. develop a user-friendly environment and to prevent or mitigate harm or destruction of computer networks,applications, devices, and data.

TOOLS USED/ALGORITHM:

- •Learning Model Algorithm
- •Naive BayesAlgorithm
- •Decision tree,
- •Support Vector Machine
- •Artificial Neural Network
- •Sequential Minimal Optimization

TECHNOLOGY:

•Machine Learning

ADVANTAGES/ DISADVANTAGES:

Advantages–Provide clear idea about the effective level of each classifier on phishing email detection.

Disadvantages-Non standard classifier

Phishing Website Detection Based on Deep Convolution Neural Network and Random Forest Ensemble Learning:

PROPOSED WORK:

It proposes an integrated phishing website detection
method based on convolution neural networks (CNN) and random forest (RF)

TOOLS USED/ALGORITHM:

•Linear Regression
•K nearest neighbor
•Support Vector Machine
•Random Forest
•XG Boost
•Naïve Bayes
•RNN Model
•CNN Model

TECHNOLOGY:

•Machine Learning , Deep Learning

ADVANTAGES/ DISADVANTAGES:

The disadvantage is that the model cannot determine whether the URL is active or not, so it is necessary to test whether the URL is active or not before detection. Advantage is that the third-party service is independent.