Detection of Phishing Websites using machine Learning Algorithm

Sneha Shirsath

MTECH-IT (Cyber Security), MIT Art, Design and Technology University, Pune
MIT School of Engineering

---***---

Abstract - *Phishing is the dishonest attempt to obtain private information from people or organisations, such as login information or account information, by posing as a reliable entity in an electronic communication. Attackers frequently use phishing techniques because it is simpler to fool a victim into clicking a malicious link that looks authentic than to try to get past a computer's security measures. Phishing attacks can put users' security and privacy at risk. This study's goal is to provide a general overview of different phishing assaults and strategies for information protection utilising machine learning-based categorization algorithms.*

Key Words: Phishing Websites, Machine Learning

1. INTRODUCTION

Phishing is a tactic used frequently to trick unwitting victims into divulging their personal information by using phoney websites. Phishing website URLs are designed to steal personal data, including user names, passwords, and online financial activities. Phishers employ websites that resemble legitimate websites both aesthetically and linguistically. Utilizing anti-phishing methods to identify phishing is necessary to stop the rapid advancement of phishing techniques as a result of advancing technology. The characteristics and machine learning-based detection methods are surveyed in this work. Attackers frequently use phishing because it is simpler to deceive someone into clicking a malicious link that appears authentic than it is to actually try to trick them.

Phishing is a type of fraud that involves presuming to be a reputable person or company in electronic correspondence while actually using a combination of social engineering to get access to sensitive and personal data, such as passwords and open-end credit information. In order to trick people into visiting fake websites using links provided on phishing websites, fake messages are produced to appear legitimate and instructed to originate from real sources such as financial institutions, online business goals, etc.

Types of Phishing Assaults: There are currently numerous distinct types of phishing attacks. These are a few of the more typical.

- Deceptive phishing

The most typical kind of phishing is deceptive phishing. In this instance, the attacker tries to get the victims to reveal sensitive information. Attackers use the data to commit financial crimes or carry out other types of offences.

- Spear phishing

Instead of focusing on a large demographic, spear phishing targets particular individuals. Attackers frequently conduct online and offline research on their targets. They can then personalise their communications and sound more genuine.

- Whaling

Whaling is the term for when attackers target a "big fish" like the CEO. These attackers frequently invest a lot of effort in profiling their victims to determine the best time and method to steal from them.

2. Related Work

This section's goal is to draw attention to other people's work that utilises various techniques to get the best results and enhance the system as a whole.

[1] Amani, Bashayr, Norah & Aram Developed a system in which they have studied 36 features out of them 3 are the new features and they have categorized them into three main categories such as Features can be extracted from URL, Features can be extracted from page content, Features can be extracted from page rank. During the study they have noticed large number of phishing website doesn't use the submit button instead they use a regular button, so we consider it as feature for phishing website and website's page that have logical structure of documents and provide accessing and manipulation for programmer to the DOM file. Afterwards, the extracted features will be sent to the classifier to produce the target label that indicates the state of the website then executes the suitable action on that.

[2] Mehmet, Our & Bane have developed the system. The proposed systems are tested with some recent datasets iand reached results are compared with the newest works in the literature. The comparison results show that the proposed systems enhance the efficiency of phishing detection and reach very good accuracy rates.

[3] Malaika, Anmol, Divyanshukumar, Gokul Developed System where they have used various algorithms Decision Tree, Random Forest classifier, K-Nearest Neihbor. Proposed

Framework executes with high efficiency, exactness and cost effectively. This framework utilizing 4 machine learning managed classification models. The four classification models have been analysed in terms of merits and demerits, performance.

[4] System has been proposed by Hauping Yuan, Xu, ukun, zhenguo, and Wenyin. The suggested approach combines URL and website link properties for the detection of phishing websites. Many classification algorithms can employ the obtained features, but DF performs competitively among them. On particular, it does not access the content of the second-level webpages and instead extracts features from the URLs and links in the first-level webpages. As a result, the suggested method works swiftly and accurately in practise. Additionally, we suggested a search operator-based approach for detecting phishing targets, which has likewise attained a respectable level of accuracy.

3. Proposed Methodology

3.1 Algorithms

The algorithms used to identify phishing websites are numerous. In this section, a few of them that can be utilised to determine whether a URL is real or phished are discussed. You can use the following algorithms to find phishing websites:

A. Artificial neural networks (ANN) A group of interconnected nodes make up an artificial neural network (ANN), which was inspired by biological neural networks (neurons). Usually, weights are assigned to each link between nodes. In the learning phase for accurate prediction, the network adjusts the weights. Because of their complicated interpretation and lengthy training periods, ANNs were viewed as being less suitable for data mining. Their benefits, however, include a high tolerance for noisy data and the capacity to classify patterns on which they have not been trained.

B. Nearest Neighbor K (k-NN) By comparing the test tuple to analogous training tuples, k-NN classifiers learn by analogy. Since these are distance-based comparisons, where each characteristic is essentially given equal weight, accuracy may suffer when noisy or irrelevant data are presented. To address the issues of pointless and noisy data tuples, respectively, editing and pruning techniques have been developed. N attributes are used to describe the training tuples. A point in an n-dimensional space is represented by each tuple. It is possible to find the ideal number of neighbours through experimentation.

C. Support Vector Machine (SVM) Both linear and nonlinear data can be classified using support vector machines (SVMs). In essence, the technique employs a nonlinear mapping to translate the original training data into a higher dimension. To keep any two classes of data apart, a linear optimum hyperplane is explored in this dimension. SVMs are also useful for numerical prediction and classification. A two-class problem with linearly separable classes is the simplest example of an SVM problem. To divide the classes in a 2-D problem, a straight line can be drawn; in fact, several lines could be drawn.

D. The Random Forests (RF) Using bagging, Random Forests can be constructed alongside random attribute selection. Random Forests use a divide-and-conquer method of performance improvement known as the ensemble approach to learning. The input or test is introduced at the top of a straightforward decision tree, and it travels down the tree, arriving at smaller subsets. The ensemble mechanism in a random forest mixes several random subsets of trees. All of the trees are explored by the input/test. The outcome is determined using the average or weighted average of the individual outcomes, or, in the case of categorical data, the majority vote. An assessment of the classifier's dependence on the strength of each individual classifier will determine how accurate a random forest is.

3.2FeatureSelection

Address bar-based features,

1. Using an IP address

If the URL, for example, uses the IP address rather than the domain name,
The user can nearly be certain that 125.98.3.123 is being used to try to steal his personal information while still leading to the necessary URL.

2. URLs with the "@" symbol

Using the "@" sign in a URL causes the browser to ignore everything before the "@," and the actual address frequently comes after the "@."

3. Redirecting using “//”

The visitor will be moved to another website if the URL path contains the character "//."

4. Adding a Prefix or Suffix to the Domain, Separated by (-)
Legitimate URLs rarely employ the dash symbol. Phishers frequently append prefixes or suffixes to the domain name, separated by (-), to give users the impression that they are visiting a trustworthy website.

A. HTML and JavaScript Based Features

1. Forwarding of websites

How frequently a website has been redirected is the thin line that separates legitimate websites from phishing ones. Customizing the Status Bar.

2. Disabling Right Click

Right-click functionality is disabled by phishers using JavaScript, preventing visitors from viewing and saving the website source code.

Exactly the same rules apply to this feature as to "Using on MouseOver to conceal the Link."

3. Using Pop-Up Windows

It is uncommon to come across a reliable website that requests customers to enter their personal information via a pop-up window.

B. Domain Based Features

1. Age of the Domain

The WHOIS database can be used to extract this characteristic. The majority of phishing websites are only active for a little time. Our analysis of our dataset reveals that a genuine domain must be at least six months old.

2. DNS Record

For phishing websites, either the claimed identity is not recognized by the WHOIS database or no records founded for the hostname. If the DNS record is empty or not found then the website is classified as "Phishing", otherwise it is classified as "Legitimate".

3. Website Traffic

By counting the number of visitors and the number of pages they view, this feature gauges the popularity of the website.

4. Page Rank

PageRank is a number between "0" and "1". PageRank attempts to gauge the significance of a website on the Internet.

5. Google Index

This function checks whether or not a webpage is indexed by Google. A website appears in search results once it has been indexed by Google.

3. CONCLUSIONS

Three key components of the study have been presented in this paper: a theory of phishing crime, a review of anti-phishing techniques from various studies, and examination of the knowledge gaps Phishing is not going away. Prior to suggesting any solutions, it is crucial to comprehend this wrongdoing. Here, we've covered a variety of phishing assault characteristics and methods for identifying malicious websites. Future work will focus on developing a phishing detection system, specifically for phishing websites, which are thought to be the most common form of assault. Instead of using a naive Bayesian strategy, we can use classifiers from the Random Forest or Artificial Neural Network. This detecting tool will assist in safeguarding users by phishing scams.

REFERENCES

[1] Malaika Rastogi, Anmol chetri, Divyanshu Kumar Singh, Rajan. Survey on detection of phishing website using Machine learning , IEEE, 2021.

[2] Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh , Dr.Aram Alsedrani Detecting Phishing Websites Using Machine Learning 78-1-7281-0108-8/19/$31.00 2019, IEEE

[3] Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri Detection of Phishing Websites by Using Machine Learning-Based URL Analysis, 2020,IEEE.

[4] Huaping Yuan, Xu Chen, Yukun Li, Zhenguo Yang, Wenyin Liu Detecting Phishing Websites and Targets Based on URLs and Webpage Links, 2018, IEEE.

[5] https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning.