Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
DOI: https://doi.org/10.17762/msea.v70i2.2447

Phishing Website Detection using Machine Learning Techniques

Bina Bhandari
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Article Info
Page Number:1577 - 1583
Publication Issue:
Vol 70 No. 2 (2021)

Article History
Article Received*: 05 September 2021*
Revised*: 09 October 2021*
Accepted*: 22 November 2021*
Publication*: 26 December 2021*

Abstract

Phishing refers to the fraudulent attempt to obtain sensitive information such as a user's username and password, as well as details about a checking account or credit card, for the purpose of using that information for malevolent purposes. It's possible that phishing scams are the most common form of cybercrime utilised today. Phishing attacks can be launched against victims in a variety of contexts, including the online payment industry, webmail, financial institutions, file hosting or cloud storage, and many more. Phishing may be detected quite effectively through the use of machine learning. Additionally, it eliminates the problem that was caused by the prior method. This study focuses on the application of machine learning technology to the problem of identifying phishing URLs. Specifically, it extracts and compares numerous characteristics of real and fraudulent URLs. By utilising the Support Vector Machine technique as well as the Random Forest technique, the project intends to identify URLs that lead to phishing websites.

Keywords –Feature Extraction, Phishing Detection, Phishing Attacks, Phishing Web-site, Machine Learning

I. Introduction

An example of a popular type of security breach known as a social engineering attack, this kind of intrusion works by convincing users that they have been tricked into divulging private information. The acquisition of sensitive information such as usernames, passwords, account numbers, and the like is the primary objective of this assault. One type of social engineering assault is known as phishing, which also goes by the name web spoofing. Phishing scams can come in many different modes of communication, including fraudulent emails, text messages, and even messaging services. Users typically have many user accounts on a variety of websites, including those for social networks, email, and even financial institutions. Therefore, the most susceptible targets for this assault are innocent online users. This is due to the fact that the majority of people are unaware of the important information that they possess, which makes it simple for this attack to be effective. The false web page is designed to look like the real web page as much as possible. As a consequence, the request made by the victim will not be sent to the victim's actual web server; rather, it will be sent to the attacker's server. In this research article, an anti-web impersonating solution is developed. The approach is built on applying machine learning to examine the URLs of bogus web pages. Using this approach, a number of steps have been constructed to check various features of websites using their Uniform Resource Locators (URLs). The Uniform Resource Locators (URLs) of fraudulent websites almost always exhibit a number of distinctive qualities that set them apart from the URLs of genuine websites.

As the number of people who use the internet rose, cybercriminals developed new methods, such as phishing, to trick victims into providing valuable information by posing as legitimate websites. This information includes account IDs, usernames, passwords, and other sensitive data. Python was used to extract features from a data set of URLs using machine learning techniques, which the suggested solution took advantage of. Using machine learning classification algorithms such as Support Vector Machine (SVM), Random Forest, and the like in order to categorise the URL as either Phishing or Legitimate.

II. Related Works

The following are some examples of works that have been published on international journals in relation to the projects:

Rao et al. proposed an innovative classification method that makes use of heuristic-based technology for feature extraction. For this, they need to classify the extracted features into the following three categories: features based on the uniform resource locator obfuscation, features based on third parties, and features based on hyperlinks. In addition, the proposed method achieves an accuracy of 99.55 percent. The disadvantage of this is that because this approach incorporates third-party features, the classification of a website's speed is sometimes dependent on the performance of third-party services. In addition to that, the success of this model is wholly contingent on the quality and quantity of the training sets. The extraction of broken links is a capability that comes with the limitation of taking a longer amount of time to execute for websites that have a greater variety of links.

The Chunlin et al. proposal suggests using a technique that focuses mostly on character frequency attributes. They have done this by combining an applied mathematics examination of the uniform resource locator with a technique called machine learning, which has led to a result that is more accurate for the classification of dangerous URLs. In addition to this, they compared six different machine-learning algorithms in order to validate the efficacy of the proposed algorithm, which provides a precision of 99.7% with a false positive rate of less than 0.4%.e effectiveness of the proposed algorithm, which provides a precision of 99.7% with a false positive rate of less than 0.4%.

The association data mining approach was utilised by Sudhanshu et al. They have proposed a rule-based classification methodology for the identification of phishing websites. They came to the conclusion that due to the association classification algorithm's straightforward rule transformation, it is superior to the other classification algorithms. They were able to reach an accuracy of 92.67% by extracting 16 features, but because this is not adequate, the proposed technique is being improved in order to get a more efficient detection rate.

A hybrid model for the classification of phishing websites was published by M. Amaadetal and colleagues. The proposed model is implemented in two stages throughout this work. During the first step, each person performs classification procedures on their own, and then they choose the three simplest models that enable high accuracy and a variety of performance criteria. In phase 2, however, they blended each individual model with the three models that gave the most accurate results to create a hybrid model that was more accurate than the individual models. On the testing dataset, they had an accuracy rating of 97.75%. The fact that it takes a longer amount of time to create hybrid models is one of the limitations of this methodology.

Phishing and non-phishing URLs can be differentiated using a feature-based method that was developed by Bhagyashree and her colleagues. This method makes use of a number of different alternatives, including lexical features, WHOIS features, Page Rank and Alexa rank, and Phish Tank-based features, in order to disguise phishing websites as legitimate ones.

III. The Proposed System

The method of machine learning is being utilised on this project. The first step in the learning process is the gathering of data using a variety of approaches and information from a variety of sources. The following stage is to prepare the data, often known as "data pre-processing," in order to resolve any problems that are caused by the data itself and to minimise the dimension of the space by removing any data that is deemed to be irrelevant (or by selecting the data that is of interest). In the context of this project, the data collection includes a significant number of URLs that are both phishing and valid. The distinction between phishing sites and legal URLs is made on the basis of the qualities that are taken from the data set. SVM, Random Forest, and Logistic Regression are the three different classification techniques that are utilised here.The models are ranked according to their performance on a number of different parameters.

A. Machine Learning

The field of study known as machine learning focuses on the creation of algorithms and methods that give computers the ability to learn and improve their intelligence based on their previous experiences. It is a subfield of artificial intelligence and has close ties to statistical analysis. Learning something literally means that the system is able to recognise and interpret the input data, which enables it to make decisions and predictions based on the data. Learning a language literally entails learning the language of another language. The first step in the learning process is compiling information through a variety of methods and obtaining it from a wide range of sources. Then, the following step is to prepare the data, also known as "pre-processing" it, in order to correct the data-related difficulties and minimise the dimensionality of the space by getting rid of the data that isn't relevant (or by selecting the data that is of relevance). It is difficult for the system to make decisions since a significant amount of data is used for learning. As a result, algorithms are built employing some statistics, probability, logic, control theory, and other relevant topics in order to synthesise the data and extract the knowledge from prior experiences. The next step is to conduct tests on the model in order to determine the reliability and functionality of the system. And last but not least, system optimisation, often known as improving the model by adding new rules or expanding the data set. Machine learning techniques can be applied in the areas of pattern recognition, classification, and prediction. Among the many areas that can benefit from machine learning are search engines, web page rankings, email filtering, face tagging and recognition, relevant marketing, character recognition, gaming, robotics, disease prediction, and traffic management.

B. Uniform Resource Locator

A web address, also known as a Uniform Resource Locator (URL), is a reference to a web resource that specifies both its location on a computer network and a method for getting it. In common parlance, a web address is also known as a URL. Many of us use the phrases "computer

address" and "Uniform Resource identifier," or "URL," interchangeably; nevertheless, a computer address is a special kind of URL.URLs are most commonly used to refer to websites (using the http protocol), but they are also utilised for database access (using the JDBC protocol), file transfer (using the ftp protocol), email (using the mail to protocol), and a wide variety of other applications. The majority of web browsers place a bar labelled "address" at the very top of a web page, where it displays the computer address of the page being viewed.

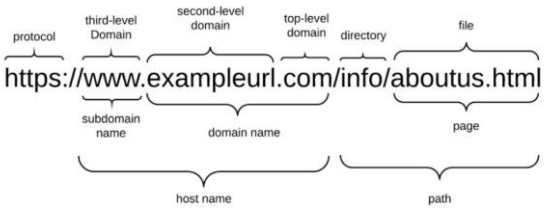

Figure 1: URLand a file name (index.html)

A typical URL is in the form http://www.example.com/ index.html, that denotes a protocol (http), a hostname (www.example.com),

C. Classification Algorithms

1. *Support Vector Machine*:

These are supervised learning methods that are utilised in order to compile data for the purposes of classification and regression. The search for the hyper plane in an N-dimensional space (N being the number of features) that correctly classifies the data points is the primary objective of the support vector machine (SVM) model. This objective requires the model to accurately categorise the data points. An SVM algorithm builds a model that assigns new examples to one of the two categories, making it a non-probabilistic binary linear classifier (even though techniques such as Platt scaling exist to use SVM in a probabilistic classification setting). Given a collection of training examples, each of which is marked as belonging to either of the two categories.

An depiction of the samples as dots in space might also be an SVM model. This illustration would need to be mapped in such a way that the examples of the various classes are separated by a distinct gap that is as large as is practically possible. After that, fresh instances are mapped into the same space, and it is predicted that they belong to a class supported by the particular facet of the hyper plane into which they fall.

It's possible that the data points that lie on each side of the hyper plane of existence belong to entirely distinct groups. A further factor that affects the size of the hyperplane is the total number of choices available. In the event where there are just two choices for the input, the hyper plane will appear as a simple line. When there are three input possibilities available, the hyper plane transforms into a two-dimensional plane. By making use of the support vectors, our goal is to increase the margin of the classifier as much as possible. It is possible for the geographical location of the hyper plane to shift after deleting the support vectors. These are the stepping stones that allow us to construct our SVM more easily.

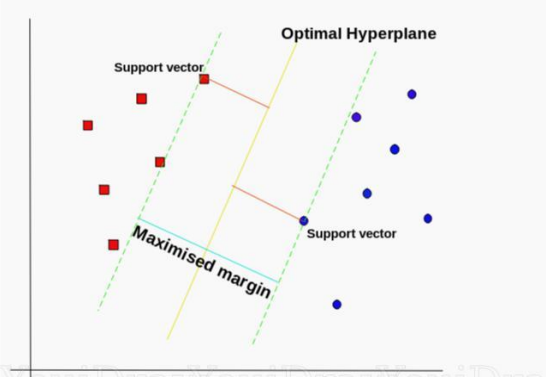

Figure 3: Support Vector Machine

2. ***Random Forest***:

A random forest is, as its name suggests, made up of a huge number of separate decision trees, each of which is responsible for making decisions independently. Each of the trees in the forest of randomness makes a classification prediction, and the classification that receives the most votes ultimately serves as the basis for our model's forecast. This is an example of an algorithm for supervised classification. The name of this algorithm comes from the fact that it generates a forest consisting of a certain number of trees. To put it succinctly, the robustness of the forest will directly correlate to the number of trees that are present in it. In a similar fashion, within the framework of the random forest classification algorithm, the high accuracy results are offered in proportion to the number of trees present in the forest.

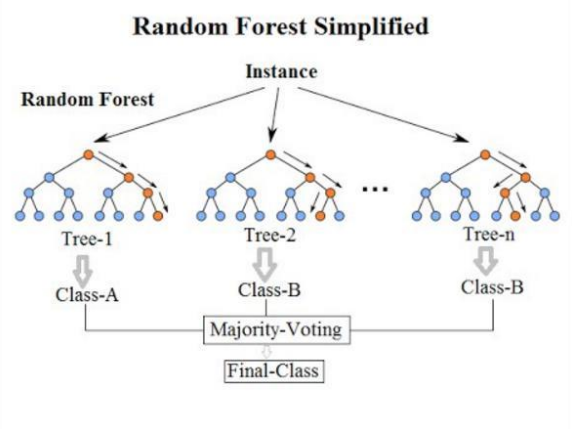

Figure 4: Random forest

- The same random forest algorithm or the random forest classifier can be used for both classification and also the regression task.
- Random forest classifier will handle the missing values.
- Random forest classifier won't overfit the model, when we have more trees in the forest,.
- The random forest classifier for categorical values can also be modelled.

IV. Conclusion

The examination of a URL for phishing purposes is highly helpful in identifying whether or not a particular URL is a legitimate URL and whether or not it should be viewed. The users are provided with a significant amount of assistance as a result in determining which websites should

not be visited. As a result, it stops them from disclosing important information to people they don't know or to sources that aren't legitimate. If you want better results, you should choose the components of the URL very carefully.In the study that we did, we carried out the classification using a robust algorithm known as Random Forest, which was developed by us. Using two distinct datasets for URL Phishing, a comparison was carried out to evaluate the relative effectiveness of linear and non-linear support vector machines (SVMs) in terms of classification accuracy. The initial dataset consisted of 31 characteristics, and it had roughly 2500 URL entries that were either phishing or not phishing. The second dataset has a total of 1353 URL entries with only 10 different attributes to choose from. On both of these datasets, Random Forest outperformed the SVM method in terms of performance. Because random forest does not suffer from the problem of overfitting, provided that the parameters are filtered and appropriately specified. As a result, it is reasonable to include them in datasets pertaining to URL phishing and to utilise them when determining whether or not a URL constitutes phishing.

It is an extremely difficult challenge to solve when trying to determine whether or not a website is authentic, often known as "phishing." Support vector machine and Random forest are the two classification models that have been utilised in the development of the model that the proposed method has utilised to detect phishing URLs. You can achieve an accuracy of 92.71% by employing all of the features in the data set in conjunction with the random forest model.When all of the accuracy metrics are taken into consideration, it is shown that random forest provides the most accurate results.

Future Enhancement

Education and awareness are the two most critical things a user can do to protect themselves from phishing attacks. Users of the Internet need to be vigilant about following all of the security advice that is supplied by specialists. In addition to this, the user should be instructed not to merely follow the links on the website to the areas where they are required to submit their personal information. Before going inside the website, it is essential to check the URL one last time. In the not-too-distant future, the system might be able to be improved such that it can automatically determine the web page and whether or not the programme is compatible with the web browser. Additional work may be done by including some extra features to differentiate between phoney web pages and authentic web pages. This can be done to distinguish between the two types of web pages. The software can also be modified to function as a web phone application, which will allow it to identify phishing on mobile platforms as well as block websites that are known to be used for phishing.

References

[1] Routhu Srinivasa Rao1, Alwyn Roshan Pais : Detection of phishing websites using an efficient feature-based machine learning framework :In Springer 2018.

[2] Chunlin Liu, Bo Lang: Finding effective classifier for malicious URL detection : In ACM,2018

[3] Sudhanshu Gautam, Kritika Rani and Bansidhar Joshi : Detecting Phishing Websites Using Rule Based Classification Algorithm: A Comparison : In Springer,2018.

[4] M. AmaadUl Haq Tahir, SohailAsghar, Ayesha Zafar, Saira Gillani: A Hybrid Model to Detect PhishingSites using Super-vised Learning Algorithms :In International Conference on Computational Science and Computational Intelligence IEEE ,2016.

[5] Ankit Kumar Jain, B. B. Gupta : Towards detection of phishing websites on client- side using machine learning based approach : In Springer Science+ Business Media, LLC, part of Springer Nature 2017

[6] Priyanka Singh, Yogendra P.S. Maravi, Sanjeev Sharma : Phish-ing Websites Detection through Supervised Learning Networks : In IEEE,2015

[7] Phishing definition, https://en.wikipedia.org/wiki/Phishing

[8] J. Han and M. Kamber, Data Mining Concepts and Techniques, Elsevier, 2006.

[9] Pradeepthi. K V and Kannan. A: Performance Study of Classification Techniques for Phishing URL Detection: In 2014 Sixth International Conference on Advanced Computing(ICoAC) IEEE,2014

[10] [Detecting Phishing Web sites: A Heuristic URL-Based Approach: In The 2013 International Conference on Advanced Technologies for Communications (ATC'13)