WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Phishing Attack Detection using Machine Learning

Artatrana Badatya

Student, BSc Information Technology, B. K. Birla College Kalyan(w), Maharashtra, India

ABSTRACT

In today world there is a significant growth of internet usage, people share their information online to connect to internet through the world. This cause a large number of loss of personal information and their transactions become more vulnerable and there is a high chance of cybercrime. Phishing is one of the cybercrime that frauds with people to get those data from them. In Phishing, mostly attackers use URLs, text and mails to steal our data. Phishing leads to serve losses of personal information, identity theft, companies and government secrets. With the help of machine learning, we detect the phishing attacks through different approaches such as Logistic Regression (LR), Classification and Regression Trees (CART), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NNet), etc. These approaches suspect website with the relevant website, if the similarity exceeds at a predefined threashold value, phishing is declared.

Keywords:- Machine learning, Logistic regression ,Cart ,SVM.

INTRODUCTION

Phishing has no universally recognised definition. However, the majority of definitions concur that the purpose of a phishing scam is to steal people's private information. Depending on how the attack is set up, the attack's media may change[1]. For instance, Pharming is a form of phishing in which the attacker tricks people into visiting phoney websites or proxy servers, generally via hijacking or poisoning the Domain Name System (DNS). In this scenario, an attacker can obtain the domain name of a target website and route all of its traffic to a phishing website without using falsified emails. Email is still the most effective method for phishing, though. The prevalence of mailers, or pre-made bulk mailing solutions, makes it easier for scammers to transmit massive amounts of fraudulent information.

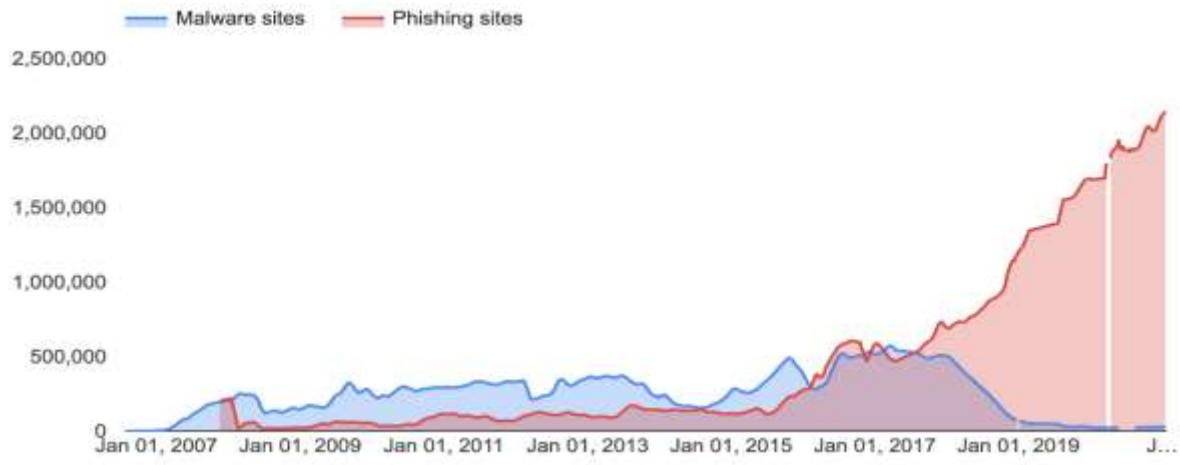

Fig 1. Increasing number of phishing attack (https://www.tessian.com/blog/phishing-statistics-2020/)

Phishing is a serious problem that is becoming worse every year. According to research conducted by Tessian in 2021, employees receive 14 phishing emails on average year[2]. Particularly hard-hit were certain sectors, with retail employees receiving an average of 49. Between May and August 2021, email-based attacks increased by 7.3%, with phishing tactics accounting for the majority of these attacks, according to ESET's 2021 research.

According to IBM research published in 2021, phishing assaults increased by 2 percentage points between 2019 and 2020, in part because to COVID-19 and supply chain uncertainties. According to CISCO's analysis on cybersecurity threat trends for 2021, 86% of firms had at least one employee who has clicked a phishing link. According to the company's research, phishing accounts for almost 90% percent

Methodology

How attackers fool user by following ways:-

- Visual Appearance : The phishing website resembles the original website. Attackers previously copied the HTML source code of an authentic website was used to create fake website.
- Address Bar : Attackers also mask the URL or address site's navigation bar using script or image. This person would they randomly believe they are entering data on the correct webpage
- Embedded objects : To avoid being detected by phishing detection methods, attackers mask the textual content and HTML coding using embedded objects (images, scripts, etc.).
- Favicon Similarity : A favicon is an image icon connected to a certain website. A website's favicon may be copied by an attacker. A phishing attempt is made if the favicon displayed in the address bar is not that of the current website.

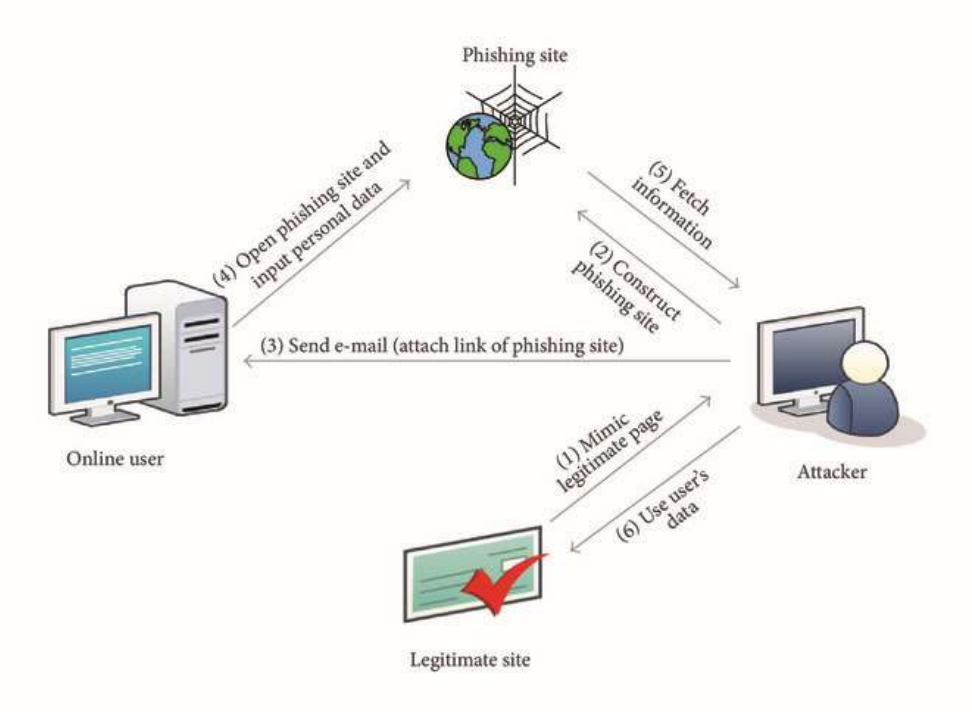

Fig 2. Phishing Mechanism (https://www.sciencedirect.com/science/article/abs/pii/S1361372321001184)

Machine learning Techniques

The majority of the machine learning algorithms covered in this article fall under the category of supervised machine learning. Here, an algorithm (classifier) makes an effort to use a certain function to translate inputs to desired outputs[3]. A classifier attempts to learn several features (variables or inputs) in classification tasks in order to predict an outcome (response). A classifier will attempt to categorise an email as phishing or legitimate (response) in the instance of phishing by learning specific characteristics (features) in the email. The next section provides an overview of two studies that use machine learning to classify phishing.An approach to categorising phishing based on the structural characteristics of phishing emails[4].

They included a total of 25 features, which included both stylistic cues (such as the phrases suspended, account, and security) and structural characteristics, such the organisation of the email's subject line and the salutation in the body.200 emails were examined (100 phishing and 100 legitimate).

They used the simulated annealing approach to choose features. They employed information gain (IG) to rank the features after selecting a feature set according to their importance[5]. Based on the chosen features, they used one-class SVM to categorise phishing emails. According to their findings, 95% of phishing emails are detected, and there are few false positives.

Phishing Detection Approaches

Logistic Regression : In many disciplines, the most popular statistical model for predicting binary data (0/1 response) is logistic regression. Its ease of use and excellent interpretability have led to its widespread application. It frequently makes use of the logit function as a component of generalised linear models. That is

$$log\frac{P(x;\beta)}{1-P(x;\beta)} = \beta^T x$$

where x is a vector of p predictors x = (x1, x2, . . . , xp), y is the binary response variable, and β is a p × 1 vector of regression parameters.

When the relationship between the data is roughly linear, logistic regression works well. However, if there are complex nonlinear interactions between the variables, it performs badly. Additionally, compared to other strategies, it requires more statistical assumptions before application. Additionally, if there are missing data in the data set, the prediction rate is impacted.

Classification and Regression Trees : The conditional distribution of y given x is described by a model called CART, or Classification and Regression Trees . Two parts make up the model: a tree T with b terminal nodes and a parameter vector = (1, 2,..., b) where I is connected to the ith terminal node. If the response y is discrete, the model can be thought of as either a classification tree or a regression tree. The predictor space is divided recursively into distinct homogenous areas using a binary tree, with the distinct regions' terminal nodes serving as the tree's nodes. Non-standard relationships can be approximated successfully by the binary tree structure (e.g. nonlinear and non-smooth).

Random forest: Each tree in a random forest depends on the values of a random vector that was randomly sampled, making them classifiers that integrate many different tree predictors. Additionally, the distribution of the forest's trees is uniform. We make the assumption that n is the number of training observations and p is the number of variables (features) in a training set in order to build a tree.

We select k p as the amount of variables to be chosen in order to identify the decision node at a tree. In order to estimate the error of the tree in the testing phase, we choose a bootstrap sample from the n observations in the training set. As a result, we randomly select k variables as a decision at a certain node in the tree and determine the appropriate split based on the k variables in the training set. Random forests are classifiers that incorporate several tree predictors. Unlike other tree algorithms, trees never get pruned; they just grow.

A data collection with several variables can be handled using random forests. Additionally, they produce an internal, unbiased estimate of the generalisation error as they are developing the forest.

They can also accurately estimate missing data. Since the method of creating the forest is random, this lack of reproducibility is a significant disadvantage of random forests. The final model has numerous independent choices trees, making it challenging to evaluate the results.

Neural Networks : The structure of a neural network is made up of numerous connected identical units (neurons). One neuron communicates with another via the interconnections. Additionally, the linkages include weights to improve neuronal delivery .Although the neurons are weak on their own, when coupled to other neurons, they are able to carry out complicated calculations. Significant interconnections play a larger role during the testing phase because weights on the interconnections are changed as the network is trained.

The Neural Network in the figure consists of one input layer, one hidden layer, and one output layer. Since interconnections do not loop back or skip other neurons, the network is called feedforward. The power of neural networks comes from the nonlinearity of the hidden neurons. In consequence, it is significant to introduce nonlinearity in the network to be able to learn complex mappings. The commonly used function in neural network research is the sigmoid function[4].

$$a(x) = \frac{1}{1+e^{-x}}$$

Support Vector Machine(SVM) :- Currently, Support Vector Machines (SVM) are among the most widely used classifiers. By increasing the distance between the closest points of the two classes, it is intended to determine the ideal separating hyperplane between them. Assume that we have two linearly separable classes with target values of +1 and -1, as well as a linear discriminating function. A selective hyperplane will fulfil:

$$w^{'} x_i + w_0 \geq 0 \text{ if } t_i = +1;$$
$$w^{'} x_i + w_0 < 0 \text{ if } t_i = -1$$

Now the distance of any point x to a hyperplane is | w 0 xi + w0 | / k w k and the distance to the origin is | w0 | / k w k. As shown in Figure 2 the points lying on the boundaries are called support vectors, and the middle of the margin is the optimal separating hyperplane that maximizes the margin of separation.

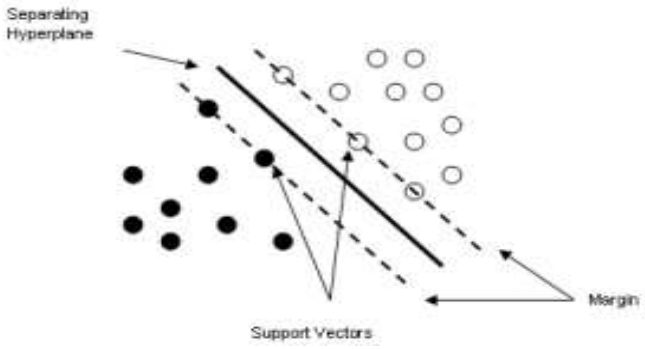

Fig. SVM

Conclusion :

In the current study we examined five different approaches by machine learning. The classifiers comprised Support Vector Machines (SVM), Random Forests (RF), Classification and Regression Trees (CART), Logistic Regression (LR), and Neural Networks (NNet). Various characteristics were learned and tested to predict phishing emails using a data set we created from raw phishing emails and many genuine emails.

The threat of phishing to web security is horrifying.The user submits personal information to a bogus website that impersonates a legitimate one in this attack. We have provided a survey on methods for visual similarity-based phishing detection. This study helps us understand phishing websites, different solutions, and the potential for phishing detection in the future. This paper discusses a variety of methods for phishing detection, although the majority of these methods still have drawbacks, such as lack of accuracy, inability to detect embedded items, inability to counteract newly discovered phishing websites, etc. These methods leverage a variety of webpage attributes, including text similarity, font colour, font size, and website graphics, to identify phishing assaults.

Acknowledgment

A special gratitude is conveyed to Prof. Swapna Augustine Nikale, Department of Information Technology of B.K. Birla College of Arts, Science and Commerce (Autonomous) Kalyan, Thane.

Reference

[1]. A. K. Jain and B. B. Gupta, “Phishing Detection: Analysis of Visual Similarity Based Approaches,” Security and Communication Networks, vol. 2017, pp. 1–20, 2017, doi: 10.1155/2017/5421046.

[2]. M. Khonji, Y. Iraqi, and A. Jones, “Phishing Detection: A Literature Survey,” IEEE Commun. Surv. Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013, doi: 10.1109/SURV.2013.032213.00009.

[3]. O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, “Machine learning based phishing detection from URLs,” Expert Systems with Applications, vol. 117, pp. 345–357, Mar. 2019, doi: 10.1016/j.eswa.2018.09.029.

[4]. N. Sanglerdsinlapachai and A. Rungsawang, “Using Domain Top-page Similarity Feature in Machine Learning-Based Web Phishing Detection,” in 2010 Third International Conference on Knowledge Discovery and Data Mining, Phuket, Jan. 2010, pp. 187–190. doi: 10.1109/WKDD.2010.108.

[5]. S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, “A comparison of machine learning techniques for phishing detection,” in Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit on - eCrime ’07, Pittsburgh, Pennsylvania, 2007, pp. 60–69. doi: 10.1145/1299015.1299021.