Phishing URL Detection using Machine Learning

Dr. R Vijayakumar, Assistant Professor / CSE
Email: msgvijayakumar@gmail.com
Sri Ramakrishna Engineering College, Coimbatore

Abishek PS	Aditya Kushwaha	Arjun RU	Ashwin Balaji PL
UG Scholar	UG Scholar	UG Scholar	UG Scholar
abishek.1901002@srec.ac.in	aditya.1901004@srec.ac.in	arjun.1901016@srec.ac.in	ashwin.1901023@srec.ac.in

Sri Ramakrishna Engineering College, Coimbatore

Abstract:

Most of the financial frauds are caused by phishing attacks. It is one of the most dangerous threats to online accounts and data, because these kinds of exploits hide behind the guise of being from a reputable company or individual and use elements of social engineering to make victims far more likely to fall for the scam.

The important thing is to exercise common sense and a good deal of caution about any message that the user receives which looks faintly suspicious and has tell-tale signs like spelling mistakes or odd phrasing, errors that malware authors often make, urges you to do something 'right now', or has a link or attachment which seems even remotely dodgy. A message which comes from a trusted source such as higher officials in the workplace or from a reputed organisation or even close friends, where their email address or details could easily have been spoofed.

The model which we have proposed is based on machine learning where the features are extracted from the input URL based on its characteristics and behaviour. The machine learning model used here is extreme gradient boosting or XGBoost, a supervised machine learning algorithm which uses more accurate approximations to find the best tree model using the training dataset to predict the URL is phishing or legitimate as target variable.

1. Introduction

The world is evolving with cutting edge technology, the people around the world are interconnected with each other through the internet with the help of electronic devices and smart gadgets. There are about 5 billion active internet users worldwide [1]. Due to the pandemic that started at the end of 2019, many traditional industries have shifted from offline mode to online. With the increase of internet usage and online services, cyber-attacks escalated around the world.

One of the majors cyberattacks faced by the people now is phishing, where the attackers use illegitimate websites to obtain victims data and use it illegally. Phishing can easily be imposed using the email and by other modes of communication. The attackers grab data easily and get valuable data. The phishing attacks lead to malware infections, loss of data, identity theft etc. The data in which these cyber criminals are interested is the crucial information of a user such as passwords, OTP, credit/ debit card details, sensitive data related to business, medical data, confidential data. Even for cautious users it's sometimes difficult to detect phishing attacks. Properly designing and deploying a Phishing URL will help block the intruders.

The attackers attempt to gain the users trust by using the same user interface, almost same URL, and cloned websites. An efficient way to detect and prevent phishing is by using an automated approach. Machine learning helps to achieve this. It is the subset of artificial intelligence used by computer systems which uses algorithms, statistical models, patterns, and inferences to complete certain tasks without human intervention. Machine learning provides the ability to automatically learn and improve from its experience without being overtly programmed. In our proposed system the algorithm goes through the process of learning with the training datasets. The major advantage of the algorithm is to allow the system to learn and decide automatically whether the website is phishing or legitimate. In terms of machine learning, a larger number of data will increase the accuracy of the model significantly. So, the prediction of phishing websites will deal with larger dataset to train the model for better accuracy.

A phishing URL and the parallel page have many features which are different from the malignant URL. The domain name of the phishing URL can be a very long and confusing name of the domain. This is very easily visible. Sometimes they use the IP address instead of using the domain name. In some cases, they can also use a shorter domain name which will not be relevant to the original legitimate website. Apart from the URL based feature of phishing detection there are many different features which can also be used for the detection of Phishing websites namely the Domain-Based Features, Page-Based Features and Content-Based Features.

2. Background and Related Works

2.1 Phishing

Phishing is a term which refers to the practice of tricking Internet users to reveal personal or confidential information. Phishing was not reported until 1996 when it was first mentioned by a popular hacker newsletter after an attack on AOL [2]. Since then, there has been an exponential increase in phishing attacks, with it becoming one of the most prevalent methods of cybercrime.

India among top 3 Asian nations affected by phishing cyber-attacks. Asia recorded an increase of 15 per cent in average cost of a phishing attack than the previous years [3]. According to the Ministry of Home Affairs, 2,00,000+ cybercrime cases have been reported in one year with more than 90% of them being financial frauds.

According to Verizon’s 2021 Data Breach Investigations Report (DBIR), phishing is the top “action variety” seen in breaches in the last year and 43% of breaches involved phishing and/or pretexting [4]. Increase in the phishing attacks compared to previous year’s reports. In 2019, phishing played a part in 78% of all Cyber-Espionage incidents and 87% of all installations of malware in the first quarter of 2019 [5]. In the earlier report by Widup (2018), it is reported that 78% of people didn’t click a single phishing link all year, meaning that 22% of the people did click one and were victims of phishing attacks. Moreover, only 17% of these phishing campaigns were reported by users. It is also emphasised that even though training can reduce the number of incidents, phish happens [6].

Since only a single message or email is needed to compromise an entire organisation, protection against it should be taken seriously. Cyber-criminals use phishing attacks to either harvest information or steal money from their victims through deceiving them with a reflection of what would seem like a regular email or website. By redirecting the victim to their disguised website, they can see everything the victim inserts in any forms, login pages or payment sites. Cyber-criminals either copy the techniques used by digital marketing experts or take advantage of the fuss created by viral events to guarantee a high click rate. Regular phishing attacks are usually deployed widely and are very generic, such that they can be deployed to target as many people as possible. A Spear Phishing attack, instead, targets a specific individual, but requires that information be gathered about the victim prior to crafting a successful spear-phishing email. A more advanced version of this attack is a Whaling attack, which specifically targets a company's senior executives to obtain higher-level access to the organisation's system. Targeted phishing attacks are increasingly gaining popularity because of their high success rates.

2.1.1 Phishing Attacks

Homograph spoofing is an attack which depends on the replacement of characters in a domain name with other visually similar characters. An example of that would be to replace 0 with o, or 1 with l. So, for a URL amazon.in the spoofed URL would be amaz0n.in. Characters from other alphabets such as Greek have also been used in the past for such attacks. The Greek o character is visually indistinguishable from the English o even though their ASCII codes are different and would redirect to different websites. Content polymorphism is addressed, as well, using visual similarity analysis of the contents.

Typosquatting targets common typographic errors in domain names. For example, an attacker could use the domain amazoon.in to target users who incorrectly type amazon.in or to trick them into clicking on a regular link [7], combat this issue using the K-Means Clustering Algorithm to observe the lexical differences between benign and malicious domains to extract features, and propose a majority voting system that takes into consideration the outputs of five different classification algorithms.

Sound squatting leverages on the use of words that sound alike (homophones) show that for a domain www.amazon.in, an adversary may use dot-omission typing errors such as wwwamazon.in, missing-character errors like www.amzon.in, character permutation errors like www.amaozn.in, character replacement errors and character insertion errors. They illustrate how they used Alexa's top one million domain list to create and register their sound squatting domains, measuring the traffic from users accidentally visiting them. Through their research, they have proven the significance of taking into account homophone confusion through abuse of text-to-speech software when tackling the issue of squatting.

Combosquatting is different from other approaches as it depends on altering the target domain by adding familiar terms inside the URLs. An example of this technique would be indianbank.com or amazon-customer-support.com. Research performed by Kintis, shows a steady increase in the use of combo squatting domains for phishing as well as other malicious activities over time [8]. It is also reported that combosquatting domains are more resilient to detection than typing error and that the majority of the combosquatting domains they were monitoring remained active for extended periods, thus suggesting that the measures set in place to counter these are inadequate.

3. Proposed Methodology

A machine learning approach predicts the URL whether it is a phishing or not. It is achieved by the extracted features from the input URL. The features are classified by most common ways of identifying phishing links. The purpose of the proposed methodology is to obtain a high accuracy in detecting phishing sites over the Internet.

3.1 Dataset

The quality of the prediction of a ML algorithm is strongly related to the quality of its training dataset. The Machine Learning approach requires a supervised learning algorithm, and therefore the samples need to be labelled as either benign or malicious. Firstly, the raw data of phishing and legitimate URLs are extracted from the official websites of phishtank and openphish, pagerank and alexa.

3.2 Feature Extraction

The URLs undergo several processes and there are around 30 characteristics of phishing websites which are used to differentiate it from legitimate ones. Each category has its own characteristics of phishing attributes and values are defined.

a) Address based features
b) Abnormal features
c) HTML and JS based features
d) Domain based features

a) Address bar based features:

i. Using IP address: If the domain of the URL of the suspected web page contains IP address, then we take it as a phishing page.

ii. Long URL to hide suspicious part: It has been a common observance that phishing web pages usually have long URLs that attempt to hide malicious URL fragments from the user. We take the assumption that a web page with a long URL is necessarily a phishing or suspicious site. In the event the assertion fails, i.e, for a legitimate web page with valid long URLs, the absence of other phishing attributes on the web page will balance the wrong assumption and correctly classify a legitimate web page as non-phishing.

iii. Use of URL shortening services: A shortened URL hides the real URL behind a redirection hop. A web page that uses a URL shortening service such as Tiny URL is highly suspicious and is likely to be a phishing attempt. Therefore, we set the rule that if the URL has been shortened using a URL shortening service then it is a phishing page and legitimate otherwise.

iv. Use of "@" symbol: Needs verification The "@" symbol is a reserved keyword according to Web standards. So the presence of "@" in a URL is suspicious and the web page is taken as phishing and legitimate otherwise.

v. Redirection with "": The presence of "//" in the URL path indicates the page will be redirected to another page. If the position of "//" in the URL is greater than seven then it is a phishing site and legitimate otherwise.

vi. Adding prefix or suffix separated by "-" to the domain: Phishers tend to add a prefix or suffix to the domain with "-" to give the resemblance of a genuine site.

vii. Sub domains and multi sub domains: If a URL has more than three dots in the domain part then it is considered as a phishing site and legitimate otherwise.

b) Abnormal based features:

viii. Request URL: A legitimate site usually has external page objects such as images, animations, files, etc. be accessed by a request URL which shares the same domain as the web page URL. We classify sites which fail this rule as phishing.

ix. URL portion of anchor tag: We check if the domain in the URL portion of all anchor tags match the main URL of the page and if the anchor tag has only URL fragments or JavaScript functions.

x. Links in <meta>, <script> and <link> tags: We check if the domain of the links in the <meta>, <script> and <link> tags matches the domain in the mail URL.

xi. Server Form Handler (SFH): When a form is submitted, some valid action must be taken. So if the action handler of a form is empty or "about:blank" or if the domain of the action URL is different from the domain of the main URL, then it is taken as a phishing site.

xii. Submitting Information to Email: If the webpage con-tains a "mailto:" function then it is taken as a phishing site and legitimate otherwise.

c) HTML and Javascript based features:

xiii. Status bar customization: Phishers can modify the status bar using JavaScript to show a legitimate URL. By analysing the "onMouseOver" events in the web page we can determine if such a modification has occurred.

xiv. Disabling right click option: Phishers can disable the right click option to prevent the user from checking the source code of the page. This is verified by analysing the source code.

xv. Using pop-up window: Legitimate sites rarely ask for user info on a pop-up window, whereas phishing sites generally use pop-up windows to get user info.

xvi. Iframe redirection: Phishers also use Iframe tags with invisible borders to get user info and redirect to the original site. We analyse the source code to check if Iframe tags are used.

d) Domain based features:

There are about 14 domain based features such as alexa pagerank, google page index, different whois results, etc.

The specified characteristics are extracted for each URL and valid ranges of inputs are identified. These values are then assigned to each phishing website risk. For each input the values range from 0, 1 or -1 based on the unique features. The phishing attributes values are represented with number -1 and 1 which indicates the attribute is present or not.

3.3 Machine Learning Model

Depending on the application and nature of the dataset used, we can use any classification algorithms. As there are different applications, we cannot differentiate which of the algorithms are superior or not. Each classifier has its own way of working and classification. After this the data is trained we shall apply a relevant machine learning algorithm to the dataset. The machine learning algorithms used are Decision Tree, Random Forest, K Nearest Neighbours Classifier, Extreme Gradient Boost, and Support Vector Machine.

I. Decision Tree: This is the most powerful and popular tool for classification and prediction. This is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

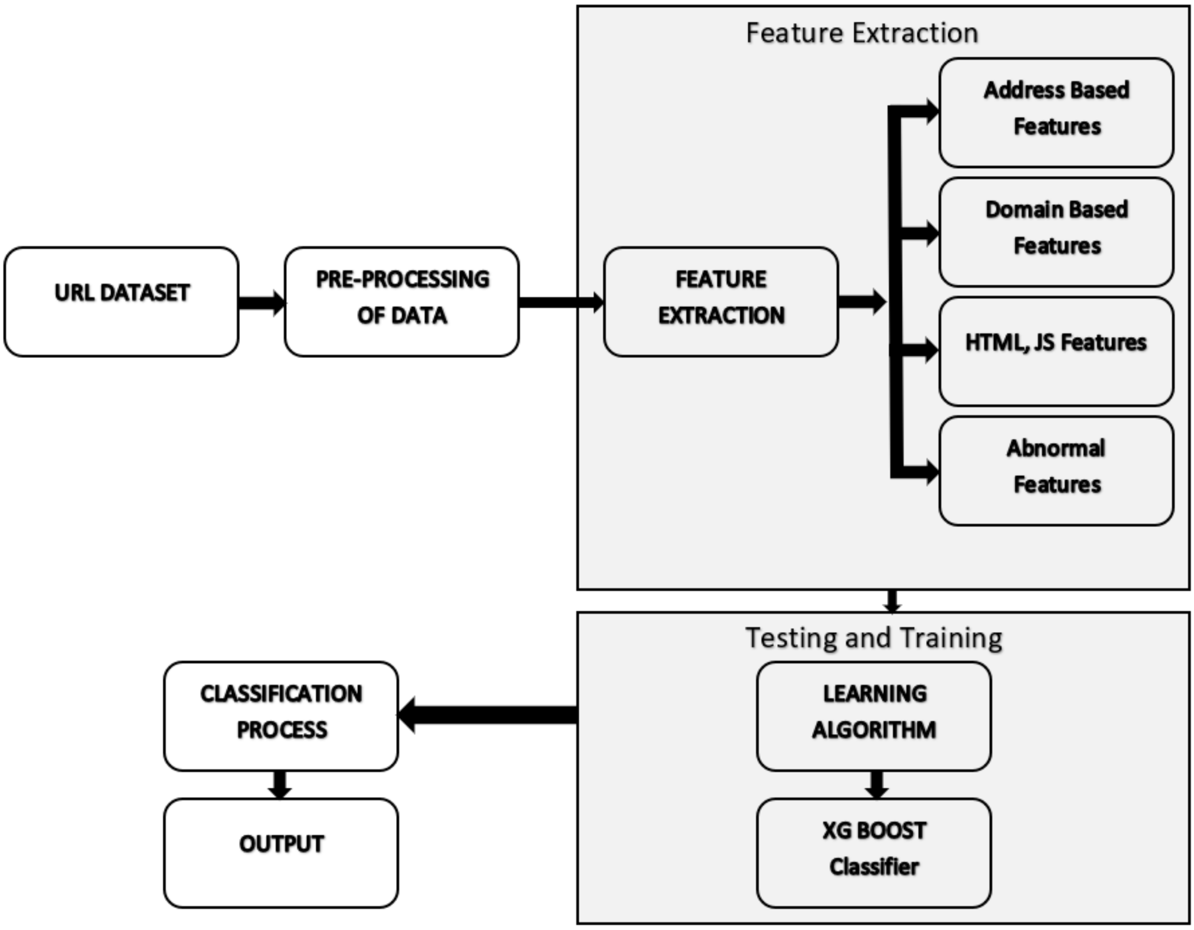
II. Random Forest: This classification algorithm is similar to ensemble learning method of classification. The regression and other tasks work by building a group of decision trees at training data level and during the output of the class, which could be the mode of classification or prediction regression for individual trees. This classifier accuracy for decision trees practice of overfitting the training data set.

III. Support vector machine (SVM): This is also one of the classification algorithms which is supervised and is easy to use. It can be used for both classification and regression applications, but it is more famous to be used in classification applications. In this algorithm each point which is a data item is plotted in a dimensional space, this space is also known as n dimensional plane, where the 'n' represents the number of features of the data. The classification is done based on the differentiation in the classes, these classes are data set points present in different planes.

IV. K-Nearest Neighbor (KNN): K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. This algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified. It can be used for Regression as well as for Classification.

V. XGBoost: Recently, the researchers have come across an algorithm "XGBoost" and its usage is very useful for machine learning classification. It is very much fast, and its performance is better as it is an execution of a boosted decision tree. This classification model is used to improve the performance of the model and to improve the speed.

These models predicted the accuracy of the detection of the phishing URL and got desired results. The testing data and evaluating the prediction with different machine learning algorithms and with the extended number of features gives more accuracy which is a bit more than other existing systems.



Fig, 3.1 Flow diagram

4. Results

We have got the desired results of testing whether the site is phishing or not by using five different classifiers. The partial implementation and the graphs of the results are given below. The project will work in such a way that when a URL is given as input to the system, it undergoes the feature extraction process specified in the Feature Extraction module. The extracted features will then go into the machine learning algorithm and predict whether the input URL is a phishing URL or a legitimate one.

The algorithm which predicts with the most accuracy is the one which we need. To find the best performing algorithm, we need the results of different algorithms. Lets see the performance of each algorithm used in the project. As this is a supervised machine learning model, the predicted values are already known so that we can find the exact accuracy of the models.

The Decision Tree algorithm shows train accuracy of 99.3% and test accuracy of 91.9%. The Random Forest algorithm which has the train accuracy of 93.5% and test accuracy of 91.9%. The K-Nearest Neighbor algorithm has the train accuracy of 99.9% and test accuracy of 88.2% while the Extreme Gradient Boosting algorithm has the train accuracy of 99.7% and test accuracy of 94.0% and The Support Vector Machine algorithm which has the train accuracy of 92.7% and test accuracy of 92.2%. The following graph 1.1 and 1.2 specifies the performance of the different algorithms.

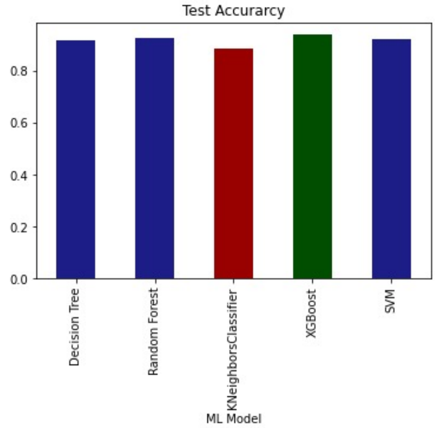


Fig. 4.1 Test accuracy

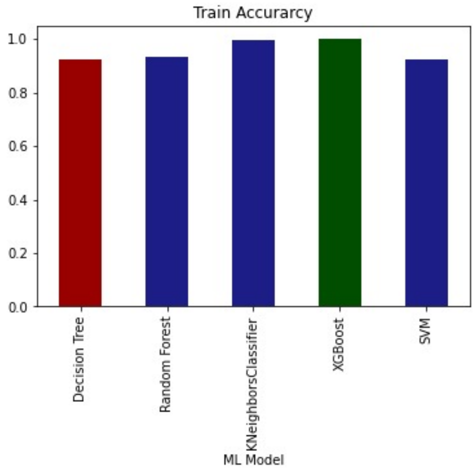


Fig. 4.2 Train accuracy

ML Model	Train Accuracy	Test Accuracy
Decision Tree	0.993	0.919
Random Forest	0.935	0.919
KNN	0.999	0.882
XG Boost	0.997	0.940
SVM	0.927	0.922

Table 4.1 Performance

5. Conclusion

Phishing is a huge threat to the security and safety of the web and phishing detection is an important problem domain. Phishing attacks are very crucial and it is important to have an automated mechanism to understand and avoid it. As very important and personal information of the user can be leaked through phishing websites, it becomes more critical to take care of this issue. This problem can be easily solved by using the machine learning algorithms with the classifier. We already have classifiers which give a good prediction rate of the phishing, but the most important thing is, the number of features. We have increased the number of features in the feature extraction part. When a right algorithm matches with the perfect model, the prediction rate will definitely be good. On reviewing some of the traditional approaches to phishing detection, such as blacklist and heuristic evaluation methods and their drawbacks, we have tested five machine learning algorithms on the dataset collected from various sources such as phishtank, openphish, etc, and reviewed their results. Based on the best algorithm and its performance a Chrome extension has been built for detecting phishing web pages directly while trying to access it. The extension allows easy deployment of our phishing detection model to end users. For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction. It can be concluded that the proposed model has proved it has high accuracy and enough features to take into action. Hopefully, phishing activities decrease and wipe out the term phishing attack from the Internet era.

6. References

[1] Joseph Johnson (2022). “Worldwide digital population as of April 2022” Published by, May 9, 2022

[2] Gunter Ollmann (2004). “Securing against the ‘threat’ of instant” Published by Network Security Volume 2004, Issue 3, March 2004, Pages 8-11.

[3] Times of India (2021). “India among top 3 Asian nations affected by DNS cyber attacks: Report”. Updated: Jun 7, 2021, 17:04 IST.

[4] Suzanne Widup, Alex Pinto, Gabriel Bassett, David Hylender (2021) Verizon Data Breach Investigations Report, May 2021.

[5] Symbol Security (author) (2019). “Verizon Data Breach Investigations Report 2019”. May 30, 2019.

[6] WidupSuzanne, WidupMarc, SpitlerDavid Hylender Gabriel (2018). “Verizon Data Breach Investigations Report 2018”. April 2018.

[7] Abdallah Moubayed, Mohammadnoor Ahmad Mohammad Injadat, Mohammadnoor Ahmad Mohammad Injadat, Ali Bou Nassif (2018). “Challenges and Research Opportunities Using Machine Learning & Data Analytics”, July 2018 IEEE Access PP(99):1-1.

[8] Kintis P, Miramirkhani N, Lever C, Chen Y, Romero-Gomez, R, Pitropakis, N, Nikiforakis, N and Antonakakis, M (2017). “Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse. Association of Computer Machinery's”, Computer and Communications Security (ACM CCS) Dallas, Texas USA 30 Oct - 02 Nov 2017.

[9] Ding, Y., Luktarhan, N., Li, K., & Slamu, W. (2019). A keyword-based combination approach for detecting phishing webpages” Computers & security, 84, 256-275.

[10] Rao, R. S., & Pais, A. R. (2019). “Jail-Phish: An improved search engine based phishing detection system”, Computers & Security, 83, 246-267.

[11] Sagar Patil, Yogesh Shetye, Nilesh Shendage (2020). “Detecting Phishing Websites Using Machine Learning”. IRJET e-ISSN: 2395-0056 Volume: 07 Issue: 02, Feb 2020.

[12] Sophiya Shikalgar , Dr. S. D. Sawarkar, Mrs.Swati Narwane (2019) – “Detection of URL based Phishing Attacks using Machine Learning “. IJERT ISSN: 2278-0181 Vol. 8 Issue 11, November-2019.