TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 5217-5220

Prediction of Phished Website at Scale Using Machine Learning

[1]G. Sai Madhukar Yadav, [1]G. Ganesh Naidu, [1]G. Charan Teja, [1]G. Devendra Kumar, [2]P. M. Mallikarjuna Shastry

[1]B Tech Students, School of C & IT, Reva University, Bengaluru, India
[2]Professor, School of C & IT, Reva University, Bengaluru, India
[1]saimadhukar1999@gmail.com, [1]ganeshnaidu343@gmail.com, [1]charan.teja158@gmail.com, [1]devad070@gmail.com, [2]mallikarjunashastry@reva.edu.in

Article Info
Volume 83
Page Number: 5217-5220
Publication Issue:
May-June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *16 May 2020*

Abstract

Phishing is one of the baiting systems utilized by phishing craftsman in the goal of misusing the individual subtleties of unsuspected clients. Identification of phishing sites is an extremely significant security measure for the vast majority of the online stages. Phishing site is a false site that seems to be comparable in appearance however changed in goal. The unsuspected clients post their information feeling that these sites originate from confided in monetary foundations. A few enemy of phishing methods rise constantly yet phishers accompany new procedure by breaking all the counter phishing components. Thus there is a requirement for effective instrument for the forecast of phishing site. Detection is done using many attributes out of this we need to identify the best set of attributes. The data set is divided into testing and training set. Further, five machine learning algorithms such as Logistic Regression, SVM(Support Vector Machine), Random Forest , Decision Tree, Neural Network have been utilized to arrange the web phishing informational index, break down the outcomes and distinguish the productive strategy to group the website page phishing informational index.

Keywords: *phishing, logistic regression, random forest, svm, decision tree, neural network.*

1. Introduction

Phishing is Fradulent system in which the assailant may endeavor to take the touchy data from the customer. This can be conceivable from multiple points of view. The assailant may send a phony email login and the customer in surge may enter his username and password. By using this touchy information they can get your Bank details and can take your money related parity. The aggressor without quite a bit of a stretch get significant data of the customers.

The assailant may mislead the customer by sending the phony sign in page which has all the earmarks of resembling the first site. There may be a slight change in the logo or there might be spelling bungles which are not seen by the customer. There may be connections like infections or key stroke, which record what you type. The top phishing assaults recently fuse back deception where hundread of bank customers got sends which guides them towards fake destinations.

In this paper we are going to divide a dataset into training and testing sets which are 80% and 20% respectively. Further we will be able to get the relative importance graph which shows the best attributes and these attributes will be processed by five machine learning algorithms.

Then Finally we will be able to get the graph which shows the reading of each and every algorithm with their respective accuracy rates.

The rest of the paper is structured as follows. Section 2 contains the previous studies done on the phishing attack. Section 3 describes the overview of phishing detection using machine learning algorithms. Section

Published by: The Mattingley Publishing Co., Inc.

May-June 2020
ISSN: 0193-4120 Page No. 5217-5220

4 presents the results and the analysis of the results. Finally, Section 5 concludes the paper.

2. Related work

Numerous analysts have recently been done in the field of phishing detection. We have assembled the data from different such works and have surveyed them which has helped us in rousing our own techniques during the time spent creation a progressively secure and exact framework.

Pawan Prakash etal.. [1] proposed a prescient blacklist approach to deal with distinguish phishing sites. It distinguished new phishing URL utilizing heuristics and by utilizing a proper coordinating calculation. Heuristics made new URL's by joining portions of the known phished sites from the accessible blacklist. The coordinating calculation at that point ascertains the score of URLs. In the event that this score is in excess of a given edge esteem it signals this site as phishing site. The score was assessed by coordinating different pieces of the URL against the URL accessible in the blacklist.

Jung Min Kang etal. [2] portrayed methodology which identified phishing dependent on clients' online exercises. This strategy kept up a white list as a piece of clients' profile. This profile was powerfully refreshed at whatever point a client visited any site. A engine utilized here distinguished a site by assessing a score and afterward contrasting it and a limit score. The score was determined from the sections accessible in the client profile and details of the present site.

Aaron Blum et al. [3] proposed a work which concentrated on the investigation of surface level highlights from URLs to prepare a certainty weighted learning calculation. The thought is to confine the wellspring of potential highlights to the character string of the URL.

The Anti-Phishing Working Group [4] distributed a contextual investigation refering to the significance of the WHO is device and how important it has been for the fast phishing site shutdown in the course of recent years all around the world.

Guang Xiang et al. [5] proposed CANTINA+, an extensive component based methodology in the writing including eight novel highlights, which abuses the HTML Document Object Model (DOM), web crawlers and outsider administrations with AI systems to recognize phish. Additionally two different channels are structured in it to help diminish FP and accomplish great runtime speedup.

Joby James et al. [6] proposed a work which with the consolidated assistance of boycotting approach and the Host based Analysis applied certain classifiers which can be utilized to help distinguish and bring down different phishing destinations. The host based, notoriety based and lexical based component extractions are applied to frame a database of highlight esteems. The database is information mined utilizing diverse AI techniques. Subsequent to assessing the classifiers, a specific classifier was chosen and was executed in MATLAB.

A. Mishra et al. [7] introduced a cross breed arrangement dependent on URL and CSS coordinating. In this methodology it can identify inserted clamor substance like a picture in a website page which is utilized to support the visual similitude in the site page. They utilized the method utilized in [3] by Jian Mao, Pei Li, Kun Li,Tao Wei, and Zhenkai Liang to look at the CSS closeness and utilized it in their strategy.

Matthew Dunlop et al. [8] proposed a program based module called goldphish to recognize phishing sites. It utilizes the site logos to recognize the phony site. The aggressor can utilize the genuine logo of the objective site to trap the web clients. Three phases to it is: Logo Extraction, Legitimate site extraction, comparision.

Ankit Kumar Jain et al. [9] recommended that visual likenesses based methods are valuable for identifying phishing sites productively. Phishing site glances fundamentally the same as in appearance to its comparing genuine site to delude clients into accepting that they are perusing the right site. Visual similitude based phishing discovery systems use the list of capabilities like content substance, content configuration, HTML labels, Cascading Style Sheet (CSS), picture, etc, to settle on the choice.

Andronicus A. Akinyelu1 et al. [10] recommended that the utilization of arbitrary woods AI calculation in order of phishing assaults, with the significant target of building up an improved phishing email classifier with better expectation precision and less quantities of highlights.

3. Proposed work

The implementation of our model is done in Five Stages which are explained below:-

1. Data Collection -The dataset is downloaded from UCI AI vault. The dataset contains 31 sections, with 30 highlights and 1 objective. The dataset has 2456 perceptions. These features are defined by W3C.Each website features in the dataset is labeled by -1 if it is not a phishing website and by 1 if it is a website used for phishing and 0 if suspicious.
2. Data Pre-Processing -Pre-Processing the data before building a model and also Extracting the features from the data based on certain conditions.
3. Training -Divide the dataset into training and testing sets as for training 80% of data and for testing 20%data.Apply machine learning algorithms like SVM, Random Forest, Decision Tree, Neural Networks, Logistic Regression etc.

A. Logistic Regression:-Fitting logistic regression and creating confusion matrix of predicted values and real values we will be able to get accuracy which was good for a logistic regression model.

B. Support Vector Machine:-Support vector machine with a rbf kernel and using gridsearchcv to predict best parameters for svm was a really good choice, and

Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 5217-5220

fitting the model with predicted best parameters we will be able to get () accuracy which is pretty good.
C. Random Forest Classification:-Random Forest are a blend of tree indicators where each tree relies upon the estimations of a subjective vector tested separately and with a similar portion for all trees in the forest. The speculation mistake for forest meets a.s. as far as possible as the measure of trees in the woods gets extraordinary. The speculation mistake of a forest of tree classifiers holds tight the quality of the individual trees in the forest and the connection between them.. Next model we tried was random forest and we will also get features importance using it, again using gridsearchcv to get best parameters and fitting best parameters to it we got very good accuracy ().Random forest was giving very good accuracy.
D. Decision tree:-Decision Tree Classification creates the yield as a binary tree like development called a decision tree. A Decision Tree model incorporates rules to anticipate the objective variable. This calculation scales well, even where there are changing quantities of preparing models and huge quantities of characteristics in huge databases.
E. Neural Network(MLP):-Multilayer Perceptron is the most much of the time utilized neural system classifier. MLP is a neural arrange and a neural system can be portrayed as a artificial neural system which comprises of an enormous number of interconnected handling segments referred to as neurons that go about as a microchip. It is a scientific model for grouping of nonlinear information into distinct classes. Multilayer Perceptron is the most famous and as often as possible utilized neural system plan. The MLP is feed forward organize engineering which includes two layers with at least one than one concealed layers; the layers are named as the info layer(relu), shrouded layer, the yield layer (sigmoid).so, the respective pickle model for every algorithm is created.pickle model(.pkl):The pickle module implements a fundamental, but powerful algorithm for serializing and de-serializing a Python object structure.pickle.dump to serialize an object hierarchy, we need to simply use dump().pickle.load to deserialize a data stream, we call the loads() function.
4.Testing or Evaluation -Apply this model we will get the respective algorithm accuracy.
. Validation -Check the variable importance graph from the results section. From variable importance graph we can tell which feature is very important. Check the accuracy plot which is also presented in the results section.rom accuracy plot we can tell which algorithm performed better for this dataset.

4. Results and Discussion

The Fig.1 explains what are the relative importance for that particular attribute. There are 30 attributes and based on these attributes we can specify whether the website is legitimate or not. So we will be only using the attributes which are having high relative importance values. Basically we use the top ten attributes which are having high relative importance value.
The Fig.2 explains which algorithm produces the maximum accuracy. The paper which we referred [11] gives around 90% accuracy we are able to achieve 96.85% accuracy using Random Forest algorithm and we are able to achieve 96.74% using SVM and 96.09 using Decision Tree. So the above graph show us the graph of the accuracy achieved by particular algorithm.

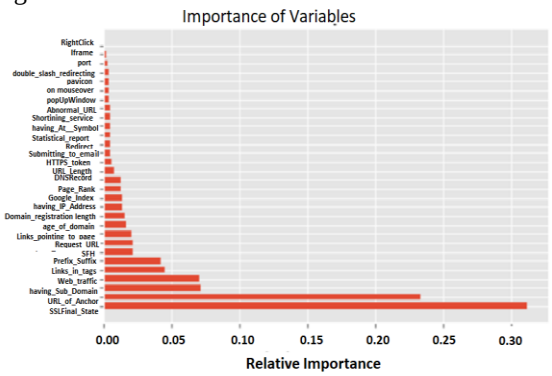

Figure 1: Importance of all the attributes

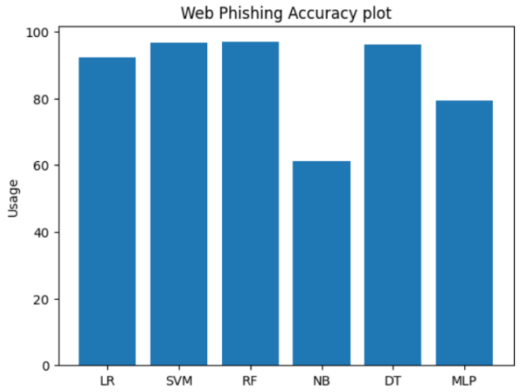

Figure 2: Accuracy levels of different algorithms

5. Conclusion

The proposed framework empowers the web clients to have a sheltered perusing and safe exchanges. Its encourages clients to spare their significant private information that ought not be spilled. Giving our proposed framework to clients as extension makes the procedure of delivering our framework a lot simpler. A specific challenge right now that crooks are continually making new methodologies to counter our guard measures. To prevail right now, need calculations that constantly adjust to new models and highlights of phishing URL's. Also, in this way we utilize web based learning algorithms. This new framework can be intended to profit greatest accuracy. Utilizing various methodologies inside and out will improve the precision of the framework, giving a productive assurance framework.. The papers which

Published by: The Mattingley Publishing Co., Inc.

May-June 2020
ISSN: 0193-4120 Page No. 5217-5220

we referred [11] gives around 90% accuracy we are able to achieve 96.85% accuracy using Random Forest algorithm.

Acknowledgement

We gratefully thank Dr. Mallikarjuna Shastry (Professor of Reva University) for guiding us to construct and implement this model and achieve the results which we expected.

References

[1] Pawan Prakash, Manish Kumar, Ramana Rao Kompella, MinaxiGupta,Purdue University, Indiana University "PhishNet: Predictive Blacklisting to Detect Phishing Attacks".

[2] JungMin Kang and DoHoon Lee "Advanced White List Approach for Preventing Access to Phishing Sites".

[3] Aaron Blum, Brad Wardman, Thamar Solorio, Gary Warner; “Lexical Feature Based Phishing URL Detection Using Online Learning”, Department of Computer and Information Sciences The University of Alabama at Birmingham, Alabama, 2016

[4] The Anti-Phishing Working Group, DNS Policy Committee;” Issues in Using DNS Whois Data for Phishing Site Take Down”,The Anti- Phishing Working Group Memorandum, 2011

[5] Guang Xiang, Jason Hong, Carolyn P. Rose, Lorrie Cranor ,”CANTINA+: A Feature-rich Machine Learning Framework for Detecting Phishing Web Sites”, School of Computer Science Carnegie Mellon University, ACM Society of computing Journal, 2015

[6] Joby James,SandhyaL,Ciza Thomas "Detection of phishing websites using Machine learning techniques", 2013 International Conference on Control Communication and Computing (ICCC)

[7] A. Mishra and B. B. Gupta, “Hybrid Solution to Detect and Filter Zero-day Phishing Attacks”, ERCICA 2014.

[8] Matthew Dunlop, Stephen Groat, and David Shelly, “GoldPhish Using Images for Content-Based Phishing analysis”, IEEE 2010.

[9] Ankit Kumar Jain1 and B. B. Gupta"Phishing Detection: Analysis of Visual Similarity Based Approaches", 2017.

[10] Andronicus A. Akinyelu1 and Aderemi O. Adewumi" Classification of Phishing Email Using Random Forest Machine Learning Technique", 2014.

[11] Ping Yi, Yuxiang Guan, Futai Zou, Yao Yao, Wei Wang, Ting Zhu,"Web Phishing Detection Using a Deep Learning Framework ".