OPEN ACCESS

Volume: 11

Special Issue: 1

Month: July

Year: 2023

E-ISSN: 2582-0397

P-ISSN: 2321-788X

Impact Factor: 3.025

Received: 14.05.2023

Accepted: 23.06.2023

Published: 01.07.2023

Citation:
Subburaj, T., and S. Diana Jennifer. "Applying Machine Learning for Identifying Fraud Sites." *Shanlax International Journal of Arts, Science and Humanities*, vol. 11, no. S1, 2023, pp. 83–88.

DOI:
https://doi.org/10.34293/sijash.v11iS1-July.6320

Applying Machine Learning for Identifying Fraud Sites

Dr. T. Subburaj
Department of Computer Applications
Raja Rajeswari College of Engineering

Diana Jennifer S
Department of Computer Applications
Raja Rajeswari College of Engineering

Abstract

Offenders looking for sensitive information create illicit clones of legitimate websites and e-mail accounts. The email will contain actual company logos and phrases. When a User selects one of these hackers' links, the hackers obtain availability of all the significant user data, including credit card detailsinformation, personal login passwords, and photos.Random Forest Decision Tree methods and are employed often in current systems, and their accuracy must be improved. The current models to be low latency. Existing systems lack a specialised user interface. Not all algorithms are compared in the present system. When consumers read the e-mails or links given, they are sent to a phoney website that looks to be from the legitimate firm. The models are used to detect phishing Websites based on URL importance factors and to discover and execute the best machine learning model. The Multinomial linear regression and other predictive modelling techniques are included in a comparison. Simple Bayes, and XG Boost. Logistic Regression beats the other two algorithms.

Keywords: Phishing Attack, Machine Learning, XG BOOST.

Introduction

Presently spoofing is a major worry for security researchers since it is not straightforward to develop a phoney website that appears to be a real website. Experts can recognise bogus websites, but not all users can, and as a result, they become victims of phishing attacks. The attacker's primary goal is to steal bank account details. Businesses in the United States lose $2 billion every year as a result of their clientele falling victim to phishing.According to the third Microsoft Computing Safer Index Report, published in February 2014, the yearly global effect of phishing might be as high as $5 billion [2]. Because of a lack of user knowledge, phishing assaults are becoming more successful. Because phishing attacks exploit user vulnerabilities, they are difficult to stop, yet it is critical to improve phishing detection systems.

Phishing is a widely used technique to mislead unsuspecting people into providing personal information by utilising phoney websites. Phishing website URLs are intended to steal individual data, such as user names, passwords, and online financial transactions. Phishers

use websites that are visually and linguistically similar to legitimate websites. To avoid the rapid evolution of phishing techniques as a result of growing technology, anti-phishing approaches must be accustomed to spot phishing. Machine learning is a practical approach for preventing phishing attacks. Hackers typically use phishing because it is easier to trick a victim into opening a malicious link that appears to be legitimate than it is to try to circumvent a computer's security mechanisms. The malicious links inside the message body are designed to seem to lead to the faked firm by using its logos and other authentic information. Machine learning is applied in the method provided to build a breakthrough way for detecting phishing websites.

Literature Survey

1. Evaluation of the Literature on Phishing Detection, this article reviews the research on phishing attack detection. Phishing attacks target holes in systems that exist owing to the involvement of humans.
2. Users are the security system's weakest link since many cyberattacks are spread via methods that benefit from of flaws in end users. Since the absence single, effective way to Describe each of the weaknesses in phishing, numerous strategies are frequently used to counteract particular attacks. Numerous the recently proposed phishing mitigation strategies are surveyed in this study.
3. Understanding and supporting users' online choices with nudges for privacy and security Information technology advancements frequently present consumers with difficult and important privacy and security choices.
4. A growing body of research has examined people's decisions when faced with privacy and information security trade-offs, the decision-making obstacles influencing those decisions, and methods to overcome those obstacles.
5. In this essay, the literature on privacy and security decision making is reviewed from a number of disciplinary perspectives.
6. It focuses on research on soft paternalistic interventions that gently steer users towards more advantageous choices to be able to support people's privacy and security decisions.
7. Warnings and preparation do not work to stop social engineering attempts,most people have faith in one another, and are open to sharing personal information. They are therefore susceptible to social engineering scams.
8. The goal of the current study was to determine how well two interventions—priming through cues to increase awareness of the risks of social engineering cyberattacks and warnings against disclosing personal information—performed in defending users against social engineering attacks.
9. Applying Spyware The website Screening K-Medoids Clustering and Probabilistic Neural Networks.The frequency and devastating effects of phishing assaults have made research into anti-phishing technologies increasingly important in information security.
10. Identifying Malicious Domain Names Created by an AlgorithmChristopher Kruegel, Giovanni Vigna, and Yong Fang. Giovanni Vigna, Christopher Kruegel, Yong Fang, Cheng Huang, Shuang Hao, and Luca Invernizzi. Gossip.

Existing Methodology

H. Huang and others, 2009 developed methods for distinguishing phishing by using a page section analogy to dissect URT tokens and produce forecasts precision phishing pages often retain their CSS style similar to their aim pages.This approach was introduced by (2017) Marchal, S., et al. to distinguish the analysis of genuine site server log information is required for phishing websites. An

off-the-shelf programme or the identification of a web phishing page. Free, demonstrates a number of exceptional qualities such as excessive precision, total autonomy, excellent linguistic freedom, swiftness of decision-making, adaptability to dynamic phishing, and adaptability to developing phishing techniques.

Mustafa Aydin et al. suggested a classification technique for recognising fraudulent websites through collecting URL characteristics from webpages and analysing subset-based feature selection methods. It uses feature extraction and selection approaches to recognise phishing websites. Alpha-numeric Character Analysis, Keyword Analysis, Security Analysis, Domain Identity Analysis, and Rank Based Analysis are the five various analyses of the extracted characteristics exist the URLs of the sites and the assembled feature matrix. The majority of these elements are textual aspects of a URL itself, while some are dependent on third-party services.

The methods of artificial intelligence that used are XG Boost, Multinomial Naive Bayes, and logarithm regression. Compared in the present system. Logistic Regression beats the other two algorithms. In the proposed system, the model is pre-processed, the words are tokenized, and stemming is conducted. The process of transforming or encoding data for simple machine transport is known as data processing. Logistic Regression has an accuracy of 96.63 percent, and the entire comparison is shown.

The current models have minimal latency. Existing systems lack a specialised user interface.The current system model is incapable of predicting a continuous result. It only works if the dependent or outcome variable is binary.If the sample size is too little, the existing system model may be inaccurate.The current situation may result in an overfitting issue.

Proposed System

We created our project with a website serving as a platform for all users. This is a responsive, interactive website that will be used to assess a web page's legitimacy real or fraudulent. This website was created utilising Lots are web design languages, including HTML, CSS, Javascript, and the Flask framework in Python. HTML is used to create the website's fundamental structure. CSS is used to enhance the appearance and usability of a website by adding effects. It should be mentioned it is as the webpage designed for all users, thus it must be simple to use and no user should have any trouble using it. Fig 1 shows the proposed architecture.

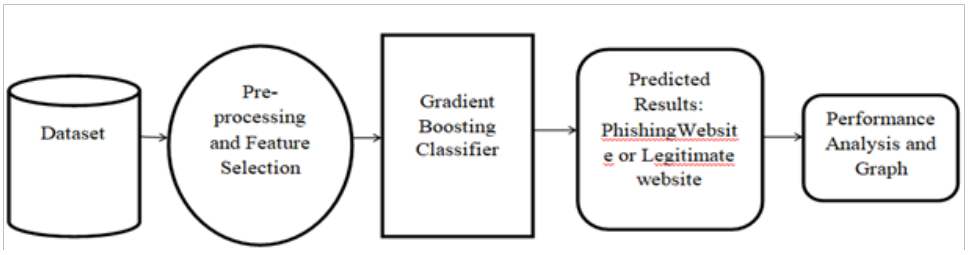

Figure 1 Proposed Architecture

The suggested system is trained using a dataset comprised of several attributes; however, the information set contains none of the website URLs. The dataset contains several criteria that must be considered while deciding if a website URL is real or fraudulent.

The Gradient Boosting Classifier applied to create the suggested system. After the equipment has undergone training using the dataset, the classifier identifies the provided URL due to the prepared information; if the location phishing, it alerts the online user phished; if the location real,

it alerts the that site is legitimate to the user. Our studies revealed fraudulent websites that a 97% accuracy using Gradient Boosting Classifier.

A user interface is given, as the example is trained using a variety of characteristics.

High Level of Precision

The suggested method is typically more accurate than previous modes, and it can train quicker, especially on bigger datasets.

Most of the suggested systems enable the processing of categorical characteristics, and several of them handle missing values natively.

The basic approach of detecting phishing websites by updating banned URLs and Internet Protocol (IP) addresses in the antivirus database, commonly known as the "blacklist" method. To avoid blacklists, attackers employ inventive tactics to deceive consumers by altering the URL to seem authentic via obfuscation and many other easy approaches such as: fast-flux, in which proxies are automatically produced to host the web-page; algorithmic production of new URLs; and so on. The main disadvantage utilising this technique is that it cannot identify zero-hour phishing attacks.

Implementation

Data Collection

The technique for gathering data is developed in the first module. This marks the beginning of the actual process of building an approach for both data mining gathering data. This step is critical because its outcome affectshow well the model works; the further as well as greater the information we gather, the better our model will work. It has numerous approaches to gathering the data, including manual interventions and online scraping. The source of the information is the well-known kaggle dataset repository.

Data Preparation

Dataset

The group of data contains 11054 unique pieces of information. Each of the 32 columns in the dataset is described below.

Gather data and get it ready for training. Remove duplicates, correct oversights, remove missing amounts, normalise, translate data types, and anything else that may require cleaning up. The effects of how we collected and/or arranged our data in a given order are eliminated by randomising the data.Create a data visualisation to assist in identifying pertinent correlations between variables or class imbalances (bias alert!) or carry out additional exploratory analyses.
Divided into training and assessment sets.

Model Selection

We utilised the method of artificial intelligence Progressive Increasing Classifier. Our training accuracy was 98.9%, thus we decided to use this approach.

Preserving the Skilled Model

The first Employing a library like Pickle, step two is to save your developed and evaluated model as an a.h5 or.pkl file. we are trustworthy enough to use it in a production-ready environment.Verify that Pickle is set up in your environment.The scenario will now be imported into the module.

Result

The results also show When more training data is utilised, the efficiency of the classifier increases. Making advantage of both the blacklist strategy and the random forest algorithm using machine learning technology, phishing websites will be more accurately identified in the future by hybrid technology.

Figure 2 Performance Analysis Page

Figure 2 shows the performance analysis of the different data

Conclusion

Using machine learning technologies, this article tries to improve detection methods for phishing websites. Using the random forest approach, we achieved 97.14% detection accuracy with the lowest false positive rate. Also, the results demonstrate filters function more effectively while more data is utilised as training data. Going forward, hybrid technology will Growing used to detect phishing websites more precisely, with the random forest algorithm using machine learning for learning technology and the blacklist approach being applied.

References

1. Gunter Ollmann, "The Phishing Guide Understanding & Preventing Phishing Attacks", IBMInternet Security Systems, 2007.
2. https://resources.infosecinstitute.com/category/enterprise/phishing/the-phishing-landscape/phishing-data-attack-statistics/#gref
3. Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013
4. Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016
5. http://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/
6. http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/
7. https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html
8. www.alexa.com
9. www.phishtank.com
10. L. Tang and Q. H. Mahmoud, "A Deep Learning-Based Framework for Phishing Website Detection", IEEE Access, vol. 10, pp. 1509-1521, 2022.
11. P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning", IEEE Access, vol. 7, pp. 15196-15209, 2019.

12. W. Ali and S. Malebary, "Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection", IEEE Access, vol. 8, pp. 116766-116780, 2020.
13. "Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks", IEEE Transactions on Network and Service Management, vol. 15, no. 3, September 2018, pp. 1076–1089; C. Pham, who was L. A. T. Nguyen, N. H. Tran, E. -N. Huh, with C. S. Hong.
14. O. Abdullateef et al., "Improving the phishing website detection using empirical analysis of Function Tree and its variants", Heliyon, vol. 7, no. 7,2021.
15. Dong-Jie Liu, Guang-Gang Geng, Xiao-Bo Jin and Wei Wang, "An effective model for detecting multistage phishing websites built on the CASE feature framework: focusing on the actual web environment", Computers Security, vol. 110, pp. 102421, 2021.