Phishing Website Detection using Machine Learning Algorithm

Dr.S.L.Jany Shabu [1], Dr.J.Refonaa [2], Dr.S.Dhamodaran [3], Dr.Vedanarayanan [4], Rishi Kumar V [5], Manoj S V [4]

[1,2,3] Associate Professor,
Sathyabama Institute of Science and Technology,
Chennai.

[4,5] Associate Professor,
Sathyabama Institute of Science and Technology,
Chennai.

Article Info
Page Number: 579- 587
Publication Issue:
Vol 71 No. 3s2 (2022)

Abstract

Phishing sites are a significant security threat. The world has spent years developing innovative methods for detecting phishing sites automatically. Even though cutting-edge solutions can boost performance, they require a lot of manual feature engineering and aren't very effective at detecting new phishing scams. The development of techniques to detect phishing websites and manage minimal phishing attempts is still a work in progress in this industry .The web page that has a lot of information that can be used to detect the maliciousness of the web server.ML is an effective technique for finding these scams. It also removes the drawbacks of the technique used before. We conducted a literature study and developed a new technique for identifying phishing websites that combines feature extraction with machine learning .The purpose of the study is to employ the information gathered to build machine learning models and deep neural networks to detect phishing sites.

Article History
***Article Received**: 28 April 2022*
***Revised**: 15 May 2022*
***Accepted**: 20 June 2022*
***Publication**: 21 July 2022*

Keywords: Phishing,Cyber Security, Machine learning, Priority based algorithms, Fusion, UCI, Python.

I. INTRODUCTION

Phishing has evolved into a major issue that affects individuals, businesses, and even entire countries. In recent years, the Web's evolution has been expedited by the availability of diverse services such as online banking, entertainment, education, software downloads, and social networking. As a result, vast amounts of data are downloaded and transferred to the Internet on a regular basis. Social engineering techniques include spoof emails that appear to be of credible firms and agencies to bring consumers to bogus sites that fool users to disclose sensitive information such as usernames and passwords Technical tactics include installing malicious software on computers in order to directly capture credentials, mechanism that often steals victims' internet account usernames and passwords.

Deceptive Phishing: Cyber criminals use a domain, or corporation to pose as a well-known organization, to gather the most secret information from victims, such as login information, passwords, bank details and so on. This sort of attack lacks depth due to the absence of personalization or modification.

Spear Phishing: The malicious URLs in this type of phishing email contain a great deal of personal information about the intended victim. An email can include details about the recipient, such as names, companies, designations, friends, coworkers and other social information.

Whale Phishing: On skewer phishing a "whale," for this situation a high ranking representative like the CEO, this sort of phishing assaults business pioneers like CEOs and high level administration laborers.

URL Phishing: An impostor or cyber-criminal uses a URL link to infect the user. People are gregarious beings who will gladly accept friend requests and might be okay to reveal personal information such as account credentials.

Because the phishers are referring users to a fake web server, this is the case. Assailants additionally utilize secure program associations to complete their criminal operations. Firms are unable to train their employees in this field because of an absence of appropriate apparatuses for forestalling phishing attacks, bringing about an expansion in phishing endeavors. As wide countermeasures, businesses are teaching their employees with simulated phishing attacks, refreshing every one of their frameworks with the most modern security processes, and scrambling delicate information. One of the most popular methods to become a victim of this phishing attack is to browse without caution. Phishing websites resemble authentic sites in appearance

Phishing is the most dangerous online criminal behavior. Since the majority of consumers use the internet to access government and financial institution services, phishing attempts have increased significantly in recent years. Phishers began earning money and have turned it into a lucrative industry. Phishers utilize a variety of tactics to assault vulnerable people, including SMS, VOIP, faked links, and malicious links.

There have been many surveys conducted by various communities all around the world to prevent phishing attacks. Phishing scams can be prevented by tracking those websites and teaching the users how to recognize them. In this study, various methods for detecting phishing websites were examined.

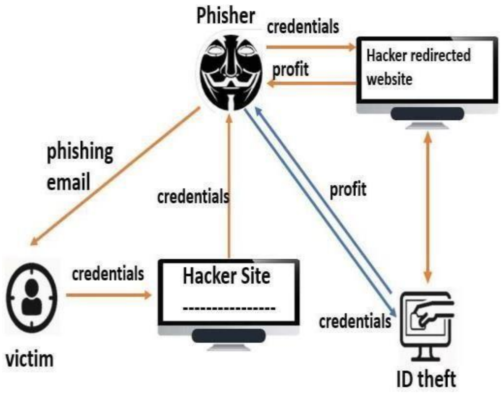


Fig. 1. Phishing attack diagram

Yong et al. fostered a clever technique for identifying phishing sites in this review that spotlights on perceiving a URL, which has been demonstrated to be a solid and productive strategy for identification. Removing shallow features from URLs is one technique. The other two, then again, make sensible element portrayals of URLs and assess URL lawfulness utilizing shallow highlights. While devouring a decent lot of time. For phishing identification, Our framework's last result is determined by joining the consequences, all things considered. Extensive testing on an Internet-source dataset shows that our technique can compete with other best detection approaches. Vahid Shahrivari and colleagues applied machine learning techniques. To detect phishing attacks, Dr.G. Ravi Kumar applied a number of ML algorithms. They employed NLP tools to improve their outcomes. Utilizing a SVM and information which was pre-handled utilizing NLP procedures, they had the option to acquire fantastic precision. Amani AlSwailem et al. For phishing detection, researchers tested a different machine learning model, but discovered that random forest was more precise. The "Fresh-Phish" open-source framework was designed by Hossein et al This framework can be utilized to make phishing site AI information. They worked with a more modest arrangement of highlights and composed the inquiry in Python. They make a tremendous, named dataset and use it to scrutinize various AI classifiers. Utilizing AI classifiers, this investigation arrives at an undeniable degree of exactness.

II. Literature Survey

Many academics have conducted research into the statistics of phishing URLs. Our method combines crucial findings from previous studies. We look back at previous work on phishing site identification using URL features, which influenced our current approach. "One of the most dangerous techniques for hackers to steal users' information such as account credentials without their knowledge," Happy says of phishing. Customers know nothing about this sort of trap, and thus, they will succumb to a Phishing trick. This could be because of an absence of monetary and individual assets, just as a deficient market mindfulness and brand trust. Mehmet et al. proposed a strategy for phishing discovery dependent on URLs in this review. The scientists utilized eight unique calculations to break down the URLs of different datasets utilizing different AI draws near and progressive constructions to think about the results. The first approach inspects the URL's numerous properties; The second method examines the site's validity by looking at where it is hosted and who manages it, while the third strategy verifies the site's actual presence. To analyze these numerous. Parts of URLs and sites, we use Machine Learning procedures and calculations. Garera et al. to characterize phishing URLs, use strategic relapse over a few significant and explicit measures. Among the elements are highlights dependent on Google's page and Google's site rank quality ideas, just as warning expressions in the URL. It's hard to make a straight examination without admittance to similar URLs and elements as our methodology.

III. Research Methodology

In this study, the linear-sequential model, commonly known as the waterfall approach, is being used. Although the waterfall method is considered traditional, it is most effective when there are negligible requirements. The application was isolated into more modest parts that were constructed utilizing systems and manually written code. Steps of this research were reading a few selected papers to evaluate the gap in this study and, as a result, defining the challenge of this study. The selection of features, classification, and detection of phishing websites were all given careful thought. The bulk of phishing detection experts rely on their own datasets, which is worth noting. However, evaluating and comparing a model's performance against that of other models is difficult since the datasets utilized were not accessible online for people who use and survey their results. As a result, such findings cannot be applied to other situations.

Language

The primary language I used to create this dissertation was Python. Python is a computer language that promotes machine learning. It comes with a number of machine learning libraries that may be used right away after being imported. Due to its enormous assortment of AI libraries, Python is generally used by engineers all around the world to manage AI .Python has a large community, thus new features are added to the latest release of the versions that’s available

Data Collection

PhishTank, an open source application, was used to collect the phishing URLs. This site offers a load of phishing URLs that is updated hourly and is available in many formats, including csv, json and a few. 5000 random phishing URLs are used to train the algorithms on this dataset.

Data Cleaning

To clear up the information, fill in the numbers left out, smooth out flung information, find then eliminate exceptions, and fix oddities.

Pre-processing is the process of transforming unstructured raw data into clean, well-organized datasets that can be used for further investigation.

Extraction of Features

There are a few procedures and information designs for phishing URL discovery in writing and business applications. A phishing URL and its associated website have a number of qualities that set them apart from malicious URLs. An attacker can, for example, establish a long and convoluted domain name to hide the genuine domain name. In the academic study detection process, different kinds of elements that are utilized in ml algorithms are utilized. Coming up next is an assortment of highlights for phishing area recognition utilizing AI moves toward that have been gotten from scholarly investigations. Some of the features may not be reasonable to utilize in certain situations due to restrictions. It may not be possible to build a fast detection algorithm capable of evaluating a large number of domains using Content-Based Features. When it comes to examining newly registered domains, page-based features aren't very useful. Therefore, the highlights that will be used by the identification instrument are dictated by the recognition component's motivation. As a result, the traits to be used in the detection technique were chosen with care.

Data is spread into thousands of training samples then some thousands of testing samples before the ML model is trained. As shown this is a supervised learning problem. The two most common forms of supervised machine learning problems are classification and regression. This data collection has a categorization challenge since the URL entered is classified as valid or phishing. The methods and models used are the following.

Algorithms

For this study, the following methods and models have been implemented.

RANDOM FOREST CLASSIFIER

This technique is one of the most commonly used in machine learning. Basically, a random forest is a mix of different decision trees. They're based on the idea that even if the individual trees are good at predicting, they almost always overfit some data. They're powerful, don't require a lot of parameter tweaking, and are not data adaptable.

MLPs

A multilayer perceptron is often referred to as a feed-forward neural network or just a neural network. MLPs undergo several phases of processing before making a final decision.

XGBoost

XGBoost stands for eXtreme Gradient Boosting. Regardless of whether regression or classification is the goal, XGBoost's gradient boosted decision trees are highly geared and extremely fast.

AUTOENCODER

A neural network with exactly as many neurons as neurons out is referred to as an auto encoder. The neural network's hidden layers have fewer neurons than input/output neurons. Because there are fewer hidden neurons, the auto-encoder must know how to encode input to them. The predictors (x) and the output (y) of an autoencoder are the same.

SVM

It is possible to classify a batch of training samples into two groups using an algorithm. During SVM training, fresh samples are assigned to one of the categories, which results in a non-probabilistic binary linear classifier.

Libraries Used

Pandas:

Python library for machine learning. Python is a free and open-source programming language. A code for bringing in and dissecting huge datasets.. Pandas is a machine learning framework that is utilized in a range of fields, counting financial aspects, money, and others. It's very easy to use, and it can show datasets in an even organization to make them more clear.

SKLEARN

Sklearn is a Python library. Sklearn comes with a number of statistical tools for classification, modeling, regression, dimension reduction, and clustering.

NUMPY

Python package Numpy is used for machine learning. The package handles arrays in Python. NumPy can be used to compute arrays of a single or two dimensions. It also contains several other functions.

MATPlotlib

MAPTlotlib is for data visualization. It's an open-source apparatus for making charts from model result. Graphics present here might help you understand the situation of the consequences. A few parts of the outcomes can be graphically shown for more straightforward understanding.

EVALUATION

A lot of machine learning models are used to assess the accuracy. The various models have been discussed in the following sections. Where the models are inspected in this review, with precision as the essential measurement. In the last stage, we thought about model exactness. The testing and preparing datasets are supported in a 20:80 proportion in all cases.

Decision Tree Classifier Experiment

To fabricate a tree, the strategy goes through all potential tests to observe the one that is generally useful with regards to the objective variable. Where we are determining the model's precision, it is dependent on information acquired from both prepared and test tests. We discovered that the test and training datasets were 82.6 percent and 81 percent accurate, respectively. Various limits are picked to make a model, and the model is then fitted in the tree. To check the model's precision, the information is partitioned into X and Y train, X and Y tests.

Random Forest Classifier Experiment

Oververfitting can be decreased by averaging the aftereffects of a few trees that all perform well and overfit in various ways. They're quite powerful, they work without a lot of parameter tweaking, and does not need data scalability. When we are forecasting the model's accuracy based on data obtained from both trained and test samples, we discovered that the accuracy were 83.4 percent and 81.4 percent, respectively.

MLP Experiment

MLPs are summed up as direct models that go through a few phases of handling prior to settling on a choice. Different boundaries are chosen to create the model, then fitted to the tree. To check model's exactness, the examples are isolated into X and Y train and X and Y test. Where we're forecasting the model's accuracy based on data obtained from both trained and test samples. The test and training datasets were found to be 86.3 percent and 85.9% accurate, respectively.

XGBoost Experiment

Different boundaries are chosen to produce the model, then fitted to the tree. To check model's precision, the examples are partitioned into X and Y train and X and Y to test. Where we're forecasting the model's accuracy based on data obtained from both trained and test samples. The test and training datasets were discovered to be 86.4 percent and 86.6 percent correct, respectively.

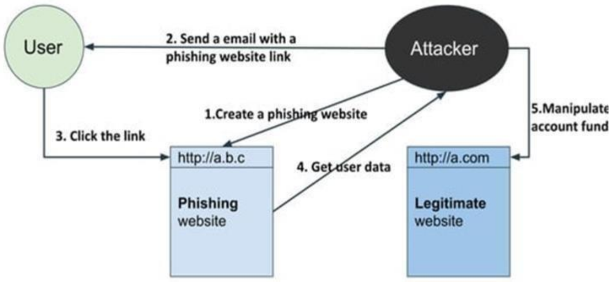
Auto encoder Experiment

The indicators (x) and result (y) of an autoencoder are similar. Various parameters are selected to generate a model, then fitted in the tree. The data is divided into X and Y trains, as well as X and Y tests, to ensure that the model is accurate. In this case, we're estimating the model's accuracy based on the data obtained from both trained and test samples. The test and training datasets were discovered to be 81.8 and 81.9 percent accurate, respectively.

SVM Experiment

SVM preparation brings about a model that orders new models into one of two groups. It changes into a non-probabilistic. A double straight classifier is given an assortment of preparing tests that are each sorted as having a place with one of the categories. Various boundaries are chosen to produce a model, then fitted in the tree. The information is isolated into X and Y trains, just as X and Y tests, to guarantee that the model is exact. When we' are determining the exactness, it is dependent on information acquired from both prepared and test tests.

R



Dataset

Phishing tank, an open-source platform, was used to get the datasets. When the data was collected, it was in csv format. The dataset contains 18 columns, and we modified it using a data pre-processing technique. To familiarize oneself with the data's features, we used a couple of the information outline techniques. Few plots and graphs are provided for visualization and to examine how the information is spread and the way in which elements are connected with each other. The Domain column has no bearing on machine learning model training. We now have 16 features and a column for the objective. In the extraction features file, the recovered features from the real and phishing URL datasets are simply concatenated, with no shuffling. To balance the distribution, we should shuffle the data and divide it into training and test datasets.

Working principle

Social engineering method that poses as authentic sites and webpages is known as phishing. The Uniform Resource Locator (URL) is the most well-known method of phishing tricks. The URL comprises file components and folders, the phisher can change it. The linear-sequential model which is commonly known as the waterfall model, was used in this study. Even so the waterfall method is popular, it works best when there are only a few factors to consider.

IV. Result and Discussion

We compared all of the machine learning models as a last step in the evaluation process. The columns of this data frame are the lists created to contain the model's results. On the training and testing datasets, accuracy of individual models we learned from our project that the XGBoost ML model has the highest accuracy compared to other models, while SVM has the least. According to the experimental data, the XGBoost strategy has the seventeenth most noteworthy worth in all of the exhibition measures assessed, showing that it is the calculation's most solid part.

This could be connected to the suggested model's overfitting avoidance strategy. To avoid overfitting, XGBOOST employs techniques such as rows sub sampling, regularization terms, shrinkage parameters, and column sub sampling. The Autoencoder has the very issue in that it takes a huge load of memory to keep the development and is deferred to execute, XGBoost, of course, has different advantages over normal point supporting techniques. These are the significant qualities of XGBoost that permit it to acquire a more prominent exactness rate than different models.

V .Conclusion & Future Work

This research compares machine learning techniques for predicting URLs. The major goal is to keep the user from obtaining critical information while maintaining security. Machine learning algorithms can be used to determine whether or not a website is genuine. We discovered that by integrating 16 characteristics, XGboost Classifier has a high accuracy when compared to other models in the research. By creating browser extensions and providing a graphical user interface, this project can be further developed. We may classify the Supplied URL as authentic or phishing using the present model.

References

[1] Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi, “Phishing Detection Using Machine Learning Techniques”, arXiv:2009.11116v1 [cs.CR] 20 Sep 2020.

[2] G. Ravi Kumar, S. Gunasekaran, and R. Nivetha R, “URL Phishing Data Analysis and Detecting Phishing Attacks Using Machine Learning In NLP,” in International Journal of Engineering Applied Sciences and Technology-2019, vol. 3, issue 10, ISSN No. 2455 -2143

[3] K. Venkateswara Rao, D. Jagan Mohan Reddy, and G. L. Vara Prasad, “An Approach for Detecting Phishing Attacks Using Machine Learning Techniques,” in Journal of Critical Reviews, vol. 7, issue 18, 2020.

[4] Amani Alswailem, Norah Alrumayh, Bashayr Alabdullah, and Aram Alsedrani, “Detecting Phishing Websites Using Machine Learning,” in International Conference on Computer Applications & Information Security (ICCAIS), 978-1-7281-0108-8/19- 2019 IEEE.

[5] Meenu, and Sunila Godara, “Phishing Detection using Machine Learning Techniques,” in International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958,vol. 9, issue 2, Dec-2019.

[6] Abdul Basit, Maham Zafar. Xuan Liu, Abdul Rehman Javed, and Zunera Jalil. Kashif Kifayat, “ A comprehensive survey of AI-enabled phishing attacks detection techniques.: Telecommunication Systems”, https://doi.org/10.1007/s11235-020-00733-2,Springer- oct,2020ISBN: 978-1-5386-0965-1.

[7] Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick, Lavanya Badiginchala, Ravali Reddy Gudur, Siri Chandana Guttha Khalilian and Nikravanshalmani, “Classification of Features for detecting Phishing Web Sites based on Machine Learning Techniques,” in International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 2, issue 8S2, June 2019.

[8] Manish Jain, Kanishk Rattan, Divya Sharma, Kriti Goel, Nidhi Gupta, “Phishing Website Detection System Using Machine Learning,” in International Research Journal of Engineering and Technology (IRJET), vol. 7, Issue 5, May-2020.

[9] S. Jagadeesan, and Chaturvedi, “URL phishing analysis using random forest”, International Journal of Pure and Applied Mathematics, vol. 118, no. 20, pp. 4159– 4163.

[10] Arun Kulkarni, and Leonard L Brown, “Phishing Websites Detection using Machine Learning”, International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, 2019.

[11] R. Kiruthiga, and D. Akila, “Phishing Websites Detection Using Machine Learning”, International Journal of Recent Technology and Engineering, vol. 8, issue 2S11, ISSN: 2277-3878, September 2019.

[12] Preeti, Rainu Nandal, and Kamaldeep Joshi, “Phishing URL Detection Using Machine Learning”, International Conference on Advanced Communication and Computational Technology, Lecturer Notes in Electrical Engineering, vol. 668, pp. 547-560, 2019.

[13] Ali Aljofey, Qiang Qu, and J.P Niyigena, “An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL”, Electronics, Electronics vol. 9, no. 1514, 2020, doi:10.3390/electronics9091514,MDPI.2020.

[14] Y. Huang, Yang, Q.J. Qin, and W. Wen, “Phishing URL Detection via CNN and Attention-Based Hierarchical RNN”, Proceedings of the IEEE International Conference On Trust, Security And Privacy in Computing And Communications, IEEE International Conference On Big Data Science & Engineering, 2019.

[15] Selvin Shabu Lilly Pushpam Jany Shabu, Kusum Yadav, Elham Kariri, Kamal Kumar Gola, Mohd AnulHaq, and Anil Kumar, “Trajectory clustering and query processing analysis framework for knowledge discovery in cloud environment”, Expert Systems. 2022; e12968. wileyonlinelibrary.com/journal/exsy.https://doi.org/10.1111/exsy.1296, 1 of 19, 2022.

[16] Ajitha, P.Sivasangari, A.Gomathi, R.M.Indira, K."Prediction of customer plan using churn analysis for telecom industry",Recent Advances in Computer Science and Communications,Volume 13, Issue 5, 2020, Pages 926-929.

[17] "Sivasangari A, Ajitha P, Rajkumar and Poonguzhali,” Emotion recognition system for autism disordered people”, Journal of Ambient Intelligence and Humanized Computing (2019)."

[18] Ajitha, P., Lavanya Chowdary, J., Joshika, K., Sivasangari, A., Gomathi, R.M., "Third Vision for Women Using Deep Learning Techniques", 4th International Conference on Computer, Communication and Signal Processing, ICCCSP 2020, 2020, 9315196

[19] Sivasangari, A., Gomathi, R.M., Ajitha, P., Anandhi (2020), Data fusion in smart transport using convolutional neural network”, Journal of Green Engineering, 2020, 10(10), pp. 8512–8523.

[20] A Sivasangari, P Ajitha, RM Gomathi, "Light weight security scheme in wireless body area sensor network using logistic chaotic scheme", International Journal of Networking and Virtual Organisations, 22(4), PP.433-444, 2020

[21] Sivasangari A, Bhowal S, Subhashini R "Secure encryption in wireless body sensor networks",Advances in Intelligent Systems and Computing, 2019, 814, pp. 679–686

[22] Sindhu K, Subhashini R, Gowri S, Vimali JS, "A Women Safety Portable Hidden camera detector and jammer", Proceedings of the 3rd International Conference on Communication and Electronics Systems, ICCES 2018, 2018, pp. 1187–1189, 8724066.

[23] Gowri, S., and J. Jabez. "Novel Methodology of Data Management in Ad Hoc Network Formulated Using Nanosensors for Detection of Industrial Pollutants." In International Conference on Computational Intelligence, Communications, and Business Analytics, pp. 206-216. Springer, Singapore, 2017.

[24] Gowri, S. and Divya, G., 2015, February. Automation of garden tools monitored using mobile application. In International Confernce on Innovation Information in Computing Technologies (pp. 1-6). IEEE.

[25] J. Refonaa, and M. Lakshmi, “Remote Sensing Based Rainfall Prediction Using Big Data Assisted Integrated Routing Framework”, Journal of Ambient Intelligence and Humanized Computing. 2021

[26] S. Dhamodaran, J. V Mahesh, Sai Swaroop, “Optimised Keyword Search With Proximity Location-Based Services”, International Conference on Computation of Power, Energy Information and Communication (ICCPEIC),ISBN: 978-1-5090-0901-5,IEEE.2017.

[27] S. Vigneshwari, B. Bharathi, T. Sasikala, and S. Mukkamala, “A study on the application of machine learning algorithms using R”, Journal of Computational and Theoretical Nanoscience, vol. 16, no. 8, pp. 3466-3472, 2019.

[28] https://archive.ics.uci.edu/ml/datasets/Phishing+Websites.